Approche multimodale pour l'évaluation d'applications de communication innovantes

Florentin Rodio

UNIVERSITÉ DE LORRAINE

L'École doctorale Stanislas

APPROCHE MULTIMODALE POUR L'EVALUATION D'APPLICATIONS DE COMMUNICATION INNOVANTES

Thèse de l'Université de Lorraine

en Ergonomie

Soutenue par Florentin RODIO

Sous la direction de Christian BASTIEN

Professeur en Ergonomie

Laboratoire PErSEUs (Psychologie Ergonomique et Sociale pour l'Expérience Utilisateurs)

Membres du jury :

Pr. BASTIEN Christian, Université de Lorraine (Directeur de thèse)
Pr. BRANGIER, Éric, Université de Lorraine (Président)
Pr. BUISINE, Stéphanie, EI.CESI Paris Nanterre (Examinatrice)
Pr. CHEVALIER, Aline, Université Toulouse Jean-Jaurès (Examinatrice)
Pr. EZZEDINE, Houcine, Université de Valenciennes (Rapporteur)
Pr. TIJUS, Charles, Université Paris 8 (Rapporteur)
GONGUET, Arnaud, Digital Innovation Project Leader, TOTAL (Invité)
MARTINOT, Olivier, Directeur de l'Innovation et des Relations Entreprises, Télécom SudParis (Invité)

Année universitaire 2016

“L’homme sage apprend de ses erreurs, l’homme plus sage apprend des erreurs des autres. ”

— Confucius

“Une science a l’âge de ses instruments de mesure parce qu’elle ne peut savoir que ce que son appareillage technique lui permet effectivement de voir”

— Gaston Bachelard

“Ce que nous désignons par l’ère du numérique se caractérise, au niveau d’observation le plus trivialement dénué d’interprétation, par un phénomène néanmoins majeur : l’irruption d’un ordinateur dans des opérations de l’ordre de la cognition, de la manipulation de données, de la connaissance, de l’information et de la communication”

— Sylvie Lele-Merviel

“Ce qui réside dans les machines, c’est de la réalité humaine, du geste humain fixé et cristallisé en structures qui fonctionnent ”

— Gilbert Simondon

REMERCIEMENTS

"La modestie est au mérite ce que les ombres sont aux figures dans un tableau: elle lui donne de la force et du relief"

— La Rochefoucault, Maximes

Mes premiers remerciements vont à Christian Bastien et Arnaud Gonguet pour leur accueil au sein du laboratoire PErSEUs de l'université de Lorraine et des Bell Labs d'Alcatel-Lucent, ainsi que pour la confiance qu'ils m'ont accordé. Je remercie également Eric Brangier pour avoir présidé le jury de cette thèse, Houcine Ezzedine et Charles Tijus pour avoir accepté d'être rapporteurs de ce travail, Stéphanie Buizine et Aline Chevalier pour leur participation en tant qu'examinateur, ainsi que Arnaud Gonguet et Olivier Martinot pour avoir participé à ma soutenance.

J'exprime également ma reconnaissance à l'Association Nationale de la Recherche et de la Technologie, qui, au travers du dispositif CIFRE, a contribué au financement de mon travail, ainsi qu'à Bruno Aidan pour m'avoir donné l'opportunité de collaborer avec les chercheurs des Bell Labs. Je tiens plus particulièrement à remercier Victor Daudonnet et Perrine De Pinel pour leurs productions réalisées durant leurs stages; Olivier Martinot, Frédérique Pain et Yann Gasté pour l'accueil qu'ils m'ont réservé au sein de leurs départements; Nicolas Marie, pour avoir travaillé ensemble autour de la recherche exploratoire; Vincent Hiribarren pour avoir élaboré les maquettes interactives utilisées dans la thèse; ainsi que Alexis Eve, Ioana Ocnarescu, Pierrick Thebault, Dominique Decotter, Cédric Mivielle, Natalie (Ebenreuter) Lehoux, Fabrice Poussiere, Jean-Baptiste Labrune pour leur bonne humeur au sein de l'équipe.

Je tiens enfin à dire un dernier mot à ma famille : MERCI. Merci de m'avoir soutenu durant ces (longues et passionnantes) années de thèse. Merci à ma mère pour ses nombreuses heures de relecture. Merci à ma famille pour leurs encouragements et les bons moments passés ensemble. Merci enfin à ma compagne et future femme pour l'amour apporté, son soutien et les moments merveilleux passés ensemble durant toutes ces années.

RESUME

Le domaine des applications de communication a connu dernièrement une grande effervescence. Parmi ses évolutions, les plus remarquables se caractérisent par la convergence des types de communication et de contenus ; un mode de gestion des applications plus communautaire et participatif; à la définition de produit, non plus seulement utile et utilisable, mais également satisfaisant une expérience utilisateur holistique ; et, enfin, suivant des modes d'interactions s'éloignant des sentiers classiques longtemps incarnés par les interfaces WIMP. La diversité des médiums émergeant, brisant le cloisonnement taxonomique précédemment établi, ne facilite pas l'application des anciennes méthodes d'évaluation basées sur les connaissances expertes et des métriques spécialisées.

Ce bouleversement a conduit les chercheurs à suivre un nouveau paradigme d'évaluation, basé sur l'expérience utilisateur. Ce courant se différencie fondamentalement de ses prédécesseurs par son caractère holistique, subjectif et positif. Un grand nombre de méthodes de mesure ont ainsi été mises au point afin de mesurer cette expérience utilisateur, mais demeurent immatures, voire contradictoires entre elles. De plus, les caractéristiques même de l'expérience utilisateur conduisent à la perte de fiabilité lors de l'exercice de sa mesure. C'est pourquoi un travail mérite d'être mené afin d'en augmenter la validité et la fiabilité.

Ce travail de thèse s'est attaché donc à utiliser plusieurs de ces méthodes, de les combiner et les articuler selon différentes techniques afin d'améliorer la qualité de la mesure. Nous nous sommes appuyés pour cela sur un large spectre d'indicateurs, d'ordre physiologiques, comportementaux et auto rapportés et de deux stratégies de triangulation en particulier : multi-facettes et multi-mesures. Enfin, ces méthodes ont testées dans des cas d'application réels et selon une complexification croissante des procédures et traitements statistiques.

Cela a donné lieu à trois études distinctes. La première a consisté à évaluer la pertinence d'un algorithme de recommandation de films face à son concurrent en utilisant une stratégie d'évaluation multi-facettes. Une deuxième étude a été élaborée afin de tester la pertinence de modèles d'évaluation multi-mesures, en évaluant l'utilisabilité de sites universitaires grâce à un logiciel de test utilisateur à distance (Evalyzer) et la combinaison multimodale de divers indicateurs d'utilisabilité. Enfin, une dernière étude a été réalisée afin de valider un protocole de mesure d'immersion multi-mesure (questionnaire, expression faciale, conductance de la peau, rythme cardiaque, comportement oculaire).

Ces trois études nous ont permis d'évaluer la pertinence d'un certain nombre de mesures (d'utilisabilité et d'expérience utilisateur), la valeur ajoutée de certaines de leurs combinaisons, ainsi qu'un retour critique sur la procédure de validation multi-facettes utilisée dans cette thèse.

MOTS CLEFS

Evaluation ; Multimodale ; Immersion ; Applications de communication ; Expérience utilisateur

ABSTRACT

The communication application domain has recently experienced some great changes, such as the convergence of communication types and content; an application management based more on participation; product definition, not only useful and usable, but also satisfying a holistic user experience; and finally, interaction modes moving away from the classic WIMP interfaces. The diversity of emerging media, breaking the previously established taxonomic partitioning, does not facilitate the application of the old assessment methods based on expert knowledge and specialized metrics.

This shift has led researchers to follow a new evaluation paradigm, based on the user experience. This trend differs fundamentally from its predecessors by its holistic, subjective and positive nature. Many measurement methods have been developed to assess the user experience, but remain immature and even contradictory. Moreover, the basic nature of the user experience has led to the loss of reliability in the exercise of its measurement. This is why studies should be conducted in order to increase the validity and reliability of this new approach.

The main aim of this research is to use several methods and to combine and articulate them using various techniques to improve the measurement quality. These studies was based on a broad spectrum of indicators (physiological, behavioral and self-reported) and two triangulation strategies in particular: multi-faceted and multi-measures. Finally, these methods were tested in real applied cases and under an increasing procedure complexity and statistical processing.

This research has resulted in three separate studies. The first aims to evaluate the relevance of a movie recommendation algorithm against its competitor using a multifaceted evaluation strategy. A second study was designed to test the relevance of a multi-measures evaluation model by assessing the usability of university sites based on a remote user testing software (Evalyzer) and a multimodal combination of various indicators of usability. A final study was conducted to validate multi-measures immersion protocol (questionnaire, facial expression, skin conductance, heart rate, eye behavior).

These three studies have highlighted the relevance of numerous measures (of usability and user experience), the added value of some of their combinations, as well as a critical return to the multi-facet validation procedure used in this thesis.

KEYWORDS

Evaluation ; Multimodal ; Immersion ; Communication Application ; User Experience

SOMMAIRE

INTRODUCTION

« *I suppose it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail.* »

Maslow (1964)

Quand j'ai intégré Alcatel-Lucent en 2010, j'ai immédiatement été fasciné par la diversité des profils disciplinaires au sein des équipes de recherche. Avoir été entouré d'ingénieurs informaticiens, de développeurs, de designers, de sociologues, et d'autres ergonomes m'a permis de prendre du recul sur ma pratique et de voir bien plus loin que les frontières habituelles de ma discipline. En effet, je suis convaincu que la diversité des approches, des méthodes, des angles de vue, constitue une richesse indispensable pour saisir toute la complexité du domaine numérique. Elle est également nécessaire pour innover, car l'innovation émerge et se nourrit de ces brassages.

Au-delà de la diversité des acteurs au sein des Bells Labs, une expression faisait toutefois consensus : l' « *Expérience Utilisateur* ». Son amélioration constituait l'horizon vers lequel tous nos efforts devaient se concentrer. Un département lui était même entièrement consacré, le : « *User Experience Design* », dont la mission principale était de concevoir ces expériences utilisateurs. Néanmoins, au-delà des préconçus, une difficulté apparaissait dès lors que l'on demandait ce qu'**était précisément** l'expérience utilisateur. A savoir, ce que désigne ce concept et comment l'appréhender dans l'optique d'une intégration dans la conception et l'évaluation des applications numériques. C'est à ce moment-là que l'on met à jour l'existence des nombreuses différences de visions, positions, tranchées, voire de chapelles de pensées, qu'engendre la discussion de ce concept. Si la diversité d'opinions est en général une force, cela prévaut seulement quand un terrain d'entente peut-être trouvé. Or, quand le sujet de discussion était l'expérience utilisateur, la disjonction des positions etait, en général, extrêmement importante. En effet, nous n'étions souvent pas très loin d'un affrontement disciplinaire, sur leurs fondations même, chacun ayant sa façon de voir le problème (prétendument « *la bonne* ») et donc sa méthode pour l'appréhender (« *correctement* »).

Cette difficulté n'est bien sûr pas spécifique aux Bells Labs d'Alcatel Lucent. Elle se retrouve en fait largement dans le monde professionnel et académique actuel. En effet, l'expérience utilisateur, ce nouveau paradigme de recherche dans le monde du numérique, a, dès sa naissance, porté la marque de la confrontation méthodologique et théorique, parfois même sous le mode de la revendication. Il faut se rappeler qu'à la fin des années 90, de nombreuses autres disciplines y ont été incorporées, suite à l'enrichissement des interfaces. Par la suite, un courant fort de contestation du paradigme dominant, celui de l'expérimentation, a commencé à émerger. De nombreux intervenants ont défendu ainsi l'utilisation d'alternatives, car, chaque méthode ayant ses forces et faiblesses, leurs croisements permettraient une augmentation de la validité générale des études. Or, s'il y a eu, depuis, une augmentation fantastique de la diversité méthodologique, peu d'efforts furent déployés pour développer les moyens de les croiser.

Pourtant, l'importance de ce dernier point, fut constamment répétée au fil du temps sans jamais avoir été toutefois véritablement exploré.

J'ai eu d'ailleurs l'occasion de constater ce paradoxe lors d'un cours sur les méthodes UX tenu par Roto et al. (2013) durant la conférence CHI en 2013. Ces derniers, parmi les plus grands spécialistes de la question, ont recensé un nombre ahurissant de méthodologies touchant à l'expérience utilisateur. Leur site « *Allaboutux.org » en* dénombre plus de quatre-vingts, contenant pêle-mêle des méthodes qualitatives, quantitatives, sociologiques, psychologiques, physiologiques, marketing, … Difficile pour un ergonome, s'intéressant de près à l'expérience utilisateur, de s'y retrouver. Ainsi, durant tout le cours, ces derniers n'ont eu de cesse de répéter l'importance de la triangulation de ces méthodes, sans vraiment expliquer ni pourquoi ni comment. Pour seul réponse, on me rétorqua que le sujet de la triangulation devait se trancher en consultant ses pairs… J'ai trouvé ce point de vue un peu naïf, car trouver un accord méthodologique entre différents spécialistes est un exercice très difficile. En effet, nous tendons à penser naturellement que nos méthodes sont supérieures à celles des autres. Maslow (1964) illustre ce fait par la loi du marteau : « *Si le seul outil que vous avez est un marteau, vous tendez à voir tout problème comme un clou* ». C'est-à-dire que nous sommes naturellement poussés à transformer la réalité d'un problème en fonction des outils dont on dispose. Pour arriver à une triangulation efficace, il faudra, tout au contraire, avoir une connaissance préalable et profonde de ces méthodes afin d'aller au-delà de leurs survalorisations respectives. Il sera ainsi possible de choisir les méthodes en fonction de leurs avantages et de les combiner pour pallier à leurs faiblesses. Pour aller encore plus loin, il faudra les retravailler pour ne garder que ce qui permet de mieux les associer, en évitant les redondances inutiles et en valorisant leurs singularités positives. Il y a donc une véritable méthodologie de la triangulation à mettre au point. Fin 2010, lors de la construction d'un laboratoire utilisateur au sein des Bells Labs ayant pour mission d'évaluer les projets produits par le département « *User Experience Design* », il m'a semblé important de travailler sur la question. En effet, si la diversité de méthodes est une richesse dans les premières phases de conception, le flou méthodologique qu'elles induisent est nuisible lors de la phase d'évaluation. Il s'agissait ainsi de prendre du recul sur les méthodes d'évaluation disponibles et de parvenir à les combiner : en somme, partir de la diversité méthodologique actuelle vers des évaluations de meilleure précision.

L'objectif de cette thèse sera donc d'explorer la piste de l'évaluation par le croisement des approches méthodologiques. L'angle précis de développement se basera sur une étude détaillée des besoins en la matière, en partant de l'état de l'art académique existant et des besoins particuliers des Bells Labs. Le cadre théorique se divise en deux parties. La première dresse un tableau général du contexte industriel de la thèse (Chapitre 1), de la recherche en méthodologie (Chapitre 2), de l'évolution des technologies de l'information et des communications (Chapitre 3 et 4), et de l'état de l'art des méthodes de conception et d'évaluation qui ont été mises au point pour assister leurs développements (Chapitre 5). La deuxième partie précise le rôle que l'ergonomie peut jouer dans ce nouveau contexte (Chapitre 6), notamment par la construction de mesures solides de l'expérience utilisateur, en se basant sur une clarification théorique de ses construits, les différentes méthodes de mesure associées, et les moyens existants pour les trianguler (Chapitre 7). Puis, trois études seront présentées dans ce manuscrit. Elles illustrent, testent et valident certaines des méthodologies d'évaluation mises au point dans la thèse, en

s’appuyant sur un traitement statistique de complexité croissante et selon le prisme méthodologique et théorique dégagé par l’état de l’art précédent. La première étude évalue la pertinence d’un algorithme de recommandations de films face à son concurrent en utilisant une stratégie d’évaluation multifacette (Théorie de la généralibilité de Cronbach). La deuxième étude évalue l’utilisabilité d’un certain nombre de sites universitaires grâce à un logiciel de test utilisateur à distance (Evalyzer) et selon un traitement statistique multifacette et de standardisation des mesures. La troisième étude a pour objectif de valider un protocole complet de mesures d’immersion multi-méthodes dans le contexte spécifique d’une activité de vidéoconférence. Enfin la thèse sera conclue par des perspectives offertes pour une éventuelle poursuite des travaux.

CONTEXTE GÉNÉRAL

CHAPITRE 1 : LA RENCONTRE FORTUITE DE L'INGENIERIE DES TELECOMMUNICATIONS ET DE LA RECHERCHE PSYCHOMETRIQUE

« We must measure what can be measured, and make measurable what cannot be measured. »

Galileo Galilei (1610)

L'âge d'or des Bells Labs et la naissance de la télécommunication moderne

L'année 1948 fut une année incroyable pour le monde des télécommunications. Elle bouleversa pour toujours les bases de notre compréhension théorique de l'information et permit le développement sans précédent d'un monde qui n'a cessé de grandir jusqu'alors : celui du numérique. La première étape fut l'invention par le laboratoire de la société téléphonique Bell d'un semi-conducteur révolutionnaire, que des ingénieurs à Muray Hills nommèrent transistor : une hybridation de **trans**conductance et var**istor** (Gleick, 2012). Cela constitua un énorme progrès face au tube électronique. C'est à partir de ce moment précis que fut réellement lancée la course à la miniaturisation technologique et le développement ubiquitaire de l'informatique. Il ne fallut pas attendre longtemps pour que ses trois inventeurs soient récompensés par un prix Nobel. Pourtant, ce n'était que le deuxième évènement le plus important de cette année, qui allait, cette fois, au-delà du seul progrès matériel.

Une monographie de 79 pages parue dans le « *Bell System Technical Journal* » de Juillet - Octobre apporta cette contribution plus profonde et fondamentale. La publication de l'article « *A Mathematical Theory of Communication »* (Shannon, 1948) marqua la naissance de la théorie de l'information moderne. Il attira tout de suite l'attention de nombreux ingénieurs, mathématiciens et d'autres scientifiques. Bien entendu, d'autres savants avaient avant lui réfléchit à la nature de l'information, mais jusqu'alors uniquement sous un prisme qualitatif. Ce que Shannon offrit de différent fut une conceptualisation mathématique de l'information. De manière comparable, Isaac Newton, en son temps, s'appropria les anciens termes de masse, force et mouvement pour en donner un sens quantitatif compatible avec une formulation mathématique. De la même manière, grâce à la quantification de l'information, Shannon permit la création d'outils et de notions précises, largement applicables à la résolution des problèmes contemporains d'ingénierie de la communication. Pour cela, une nouvelle unité de mesure fut inventée : le Binary digIT, le « bit ». La purification du concept d'information une fois faite, rendue simple, distillée, comptée en bits, nous priment conscience que l'information se trouvait

partout (Gleick, 2012). Elle présente un enjeu vital pour l'organisme vivant même le plus simple et est un outil indispensable à maîtriser pour faire évoluer nos systèmes d'information artificiels. L'idée de transmettre des données, peu importe le contenu, à l'aide d'un flux de 0 et de 1, était née. 23 autres théorèmes issus de cette théorie permirent de représenter efficacement l'information, la coder, la compresser, la chiffrer, la contrôler, la transmettre en présence de perturbation, etc. La radiotéléphonie numérique et les CD-Rom n'ont pu être mise au point qu'à partir des procédés de codage efficaces issus de la théorie de l'information.

Un contexte riche et favorable

Il est certain que l'inspiration de Shannon s'est nourrie de tous les environnements dans lesquels il a pu étudier et travailler. Il étudia en premier le génie électrique et les mathématiques à l'université du Michigan avant de passer sa maîtrise au MIT en 1938. Puis, pendant la guerre, il sera chargé d'élaborer un système de tir anti-aérien se basant automatiquement sur les données radars. Son premier rapport, *« lissage des données et prédiction dans les systèmes de conduite des tirs »* pointe déjà vers une généralisation du traitement du signal. Il poursuit son travail pour les services secrets de l'armée américaine en se concentrant sur les fondements théoriques de la cryptographie. Il sera chargé, entre autres, de déceler les messages cachés par brouillage dans le code ennemi et de sécuriser la ligne de communication entre le président Roosevelt et Winston Churchill (système X).

C'est en 1941 que Shannon rejoint les Bells Labs, où il fut recruté pour améliorer les méthodes de transmission de l'information. Durant cette période, il n'eut pas de compte à rendre, même si sa hiérarchie ne comprenait pas exactement ce sur quoi il travaillait. En effet, à cette époque, les laboratoires d'AT&T n'attendaient pas une monétisation immédiate du travail de leurs chercheurs, qui pouvaient faire des détours en mathématique ou en astrophysique (Gleick, 2012). De ce fait, de nombreux profils de mathématiciens et d'ingénieurs se côtoyaient. Malgré tout, le sujet central de la compagnie de téléphone restait assez peu exploré, bien que plus de 125 millions de conversations transitaient déjà par les systèmes de l'entreprise.

C'est grâce à cette liberté, l'expérience théorique acquise et la diversité des sujets au sein des Bells Labs gravitant autour de la transmission des informations (téléphoniques, télégraphiques, radiophoniques, …) que Shannon fut capable de lier ces différents domaines dans un seul cadre en créant une théorie générale de l'information.

La diffusion de la théorie de l'information au-delà des télécommunications

L'une des caractéristiques fondamentales de la théorie −et contre intuitive à première vue− est l'exclusion de la sémantique. La théorie de l'information est, en effet, indifférente à la signification des messages, que Shannon préférait laisser aux philosophes. En effet, pour l'ingénieur en communication, l'information que porte un message n'a pas d'incidence sur les moyens pour la transporter. Néanmoins, l'information de Shannon n'est qu'une vision existante sur ce qu'est l'information. En résumant la littérature technique, on peut décrire trois grandes

postures qui décrivent ce qu'est l'information (Adriaans & van Benthem, 2008a). L'information de type A, qualitative, est liée à la connaissance, la logique épistémique[1] et la sémantique[2]. Elle traite de ce qui est transmis dans un message informatif, soit exactement ce que Shannon a délibérément exclu. L'information de type B est probabiliste et quantitative. Elle est liée au concept d'entropie en physique et directement à la théorie de l'information de Shannon. Enfin, l'information de type C est algorithmique et quantitative. Elle est liée aux travaux de Kolmogorov et aux fondations de la programmation. On voit bien ici que le champ de la théorie de Shannon sur l'information est parcellaire, limité et ne traite qu'une partie de ce que l'on entend par « information ». Pourtant, dans les années 50, c'est la théorie qui aura le plus d'impact sur les autres domaines scientifiques de l'époque.

En effet, la théorie construit un pont entre les mathématiques et l'ingénierie électrique, puis fonde en grande partie l'informatique, dont l'étymologie démontre l'origine théorique. De même, la majeure partie des théorèmes statistiques peuvent être reformulée selon les principes d'entropie et de divergence d'informations. Mais le plus grand apport de la théorie, c'est de nous ouvrir les yeux sur ce principe fondamental qu'est l'information. Michel Serres résuma cela très bien lors d'une communication orale organisée lors des 40 ans de l'INRIA à Lille, en 2007 : « *[...] Je ne connais pas d'être vivant, cellule, tissu, organe, individu et peut-être même espèce, dont on ne puisse pas dire qu'il stocke de l'information, ou qu'il traite de l'information, qu'il émet, qu'il reçoit de l'information. Cette quadruple caractéristique est si propre au vivant que l'on serait tenté de définir la vie de cette manière. Mais nous ne pouvons pas le faire parce que les contre-exemples surabondent, en effet je ne connais pas d'objet du monde, atome cristal, montagne, planète, étoile, galaxie, dont nous ne puissions pas dire de nouveau qu'il stocke de l'information, qu'il traite, qu'il émet, qu'il reçoit de l'information. Cette quadruple caractéristique est donc commune à tous les objets du monde, vivants ou inertes. Nos sciences dures qui ne parlaient autrefois que de force et d'énergie parlent depuis assez récemment de code et ce que l'on appelle le « doux ». Les sciences dures s'occupent aussi du doux. Cela dit, je ne connais pas non plus d'association humaine, une famille, une ferme, un village, une métropole, une nation, dont on ne puisse pas dire de nouveau qu'elle stocke de l'information, qu'elle traite, qu'elle émet, qu'elle reçoit de l'information. Voici donc une caractéristique commune aux sciences humaines et aux sciences dures, c'est-à-dire à notre existence et à notre environnement, de telle sorte que le jour où nous avons inventé un objet qui stocke, traite, émet et reçoit de l'information – je veux parler de l'ordinateur – nous avons inventé un outil qui peut s'appeler universel. Pourquoi ? Parce qu'il mime d'une certaine façon le comportement des choses dont je viens de parler. Quelles choses, mais toutes choses ! D'où le caractère universel de cet outil, non pas seulement parce qu'il peut servir à tout, mais parce qu'il mime le comportement, la conduite, le profil des choses de ce monde. Changement, révolution culturelle ou cognitive, mais premièrement pratique [...]* ».

En biologie, le concept d'information génétique a permis un développement sans précédent de la biologie moléculaire (Keller, 2000; Kay, 2000; Yockey, 1992). À partir de ce moment, la

[1] Branche de la logique non classique s'intéressant à la formalisation d'un système ayant pour objectif de décrire de manière satisfaisante les propriétés et les relations mutuelles des notions de savoir et de croyance

[2] Branche de la linguistique qui étudie le sens des énoncés

biologie devint une science de l'information totale. En effet, on peut constater que les gènes encapsulent de l'information que l'on peut lire et recopier ; les cellules d'un organisme sont autant de nœuds dans un réseau de communication richement interconnecté, recevant et transmettant, codant et décodant de l'information ; l'évolution elle-même incorporant en ses fondements un échange incessant d'informations entre organismes et environnement ; et l'ADN constitue la plus aboutie des technologies de stockage et de traitement d'informations de l'univers connu à un niveau cellulaire : un alphabet et un code de 6 milliards de bits pour un être humain (Gleick, 2012). Le célèbre théoricien de l'évolution Richard Dawkins le confirme : « *Ce qui est à la source de la vie, ce n'est pas un feu, pas un souffle chaud, pas une «étincelle de vie», c'est de l'information, des mots, des instructions… Si vous voulez comprendre la vie, ne pensez pas à une bouillie ou à un gel vibrant et palpitant, pensez en termes de technologie d'information* » (Dawkins, 1989, p. 12).

Il n'a pas fallu non plus attendre longtemps avant que les linguistes et les psychologues s'approprient également cette théorie révolutionnaire. Pour les linguistes, ce fut simple et rapide, car Shannon lui-même chercha à illustrer sa théorie de l'information à partir des caractéristiques du langage humain. Par exemple, il chercha à calculer l'entropie de la langue anglaise et estima sa redondance à 69% (Shannon, 1951). Cela explique que l'on puisse comprendre assez facilement une phrase même en l'absence de quelques lettres, mots ou en présence de bruit. Jakobson fut un des linguistes les plus enthousiasmé par la théorie de Shannon, et s'en inspira grandement pour mettre au point son modèle décrivant les différentes fonctions du langage (Jakobson, 1963).

Les débuts des sciences cognitives ont également été largement inspirés par la théorie de l'information. De nombreux psychologues expérimentaux du milieu du siècle dernier ont été attirés par la capacité à mesurer les aptitudes et limitations cognitives (Boden, 2008) à partir de ce modèle. Les premiers exemples de psychologue inspirés par cette théorie incluent Herbert Simon, Donald Broadbent, George Miller et Jérôme Bruner. Ce sont toutes des figures importantes du développement des sciences cognitives à leurs débuts. Herbert Simon étudia les processus de décision en groupe et suggéra à travers sa théorie de la rationalité limitée que « *chaque organisme humain vit dans un environnement qui produit des millions de bits de nouvelles informations chaque seconde, mais le goulot d'étranglement de l'appareil de perception n'admet certainement pas plus de 1000 bits par seconde et probablement moins* » (Simon, 1959, p. 273) ; de ce fait, la raison est limitée car fonctionnant à partir d'informations incomplètes. Broadbent (1958) étudie la perception, la communication, et l'attention chez l'homme. Il met au point sa théorie du filtre attentionnel, selon laquelle l'attention joue le rôle d'un filtre de bas niveau qui sélectionne une partie de l'information qui parvient du monde extérieur à la périphérie sensorielle. Miller (1956) met en évidence la limite de notre mémoire à court terme et le besoin de segmenter les informations pour mieux les retenir. Il est célèbre pour le nombre magique sept, plus ou moins deux, qui correspond en moyenne au nombre d'objets pouvant tenir dans la mémoire de travail d'un humain. Enfin, pour Bruner, la théorie de l'information est une source clé pour ses expériences célèbres dans le domaine de l'apprentissage de l'abstraction (Bruner, Goodnow, & Austin, 1956). Pour les psychologues cognitivistes, le cerveau est considéré comme un système complexe de traitement de l'information. Ce système fonctionne grâce à des structures de stockages (la mémoire) et à des

opérations d'analyse logique (comme la recherche en mémoire ou l'identification de catégories). Le rapprochement entre ces deux entités, le cerveau et l'ordinateur, sera particulièrement fécond et permettra d'apporter un changement radical dans notre compréhension de l'humain et des mécanismes d'apprentissage.

Néanmoins, les psychologues jusqu'alors percevaient l'information sous le seul prisme du type A, c'est-à-dire celui porteur de sens. Il a fallu supplanter la notion ordinaire d'information par celle de Shannon, − c'est-à-dire basée sur le calcul logique et probabiliste− pour que naisse véritablement le domaine des sciences cognitives (Boden, 2008). C'est grâce à la réflexion conjointe de neurologues, psychologues cogniticiens, informaticiens et mathématiciens que l'on a pu montrer explicitement que tout le champ de la psychologie n'est qu'un problème de *« formalisation de systèmes de calcul interconnectés capable d'implémenter les différents phénomènes psychologiques* » (Boden, 2008). Dès lors, les sciences étudiant les mécanismes de la pensée humaine et artificielle n'ont cessé de se nourrir l'un l'autre. On sait que l'architecture de la quasi-totalité des ordinateurs modernes se base sur celle de Von Neumann, un célèbre mathématicien qui s'est inspiré pour cela des fonctions du cerveau humain et des opérations logiques des neurones, telles que décrits dans les travaux de Warren Mc Culloc et Walter Pitts (1943). Très récemment, l'inspiration a été puissée dans le sens inverse, des modèles mathématiques vers les sciences naturelles. En effet, de nombreux neuropsychologues se sont rendus compte que certains modèles probabilistes bayésiens utilisés en informatique permettent d'expliquer les mécanismes d'apprentissage chez l'être humain avec une efficacité incroyable. Ainsi, l'hypothèse du « *cerveau bayésien* » (Doya, 2007) se révèle particulièrement productive dans de nombreux champs des neurosciences cognitives actuelles. En effet, l'étude du comportement animal et humain sous ce prisme suggère que les adultes et les enfants possèdent une vaste capacité d'inférence statistique à de multiples niveaux (perception, action, langage…). L'architecture même du cortex pourrait s'expliquer par la réplication, à plusieurs niveaux hiérarchiques, d'un même circuit neuronal bayésien (Dehaene, 2012). Ces inférences probabilistes sont accessibles à des enfants de quelques mois (Xu & Garcia, 2008) et sont à la base des expériences d'apprentissage des mots chez l'enfant entre 17 et 19 mois (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). Nous pouvons en conclure, comme Paul Valéry, que « *nous comprenons d'autant mieux les vivants que nous inventons et construisons des machines* » (1959, p. 617).

La rencontre avec l'épistémologie et la psychométrie

Enfin, un lien fort existe entre la théorie de l'information de Shannon et l'épistémologie. C'est un lien de première importance pour nous, car cette thèse puise directement dans cette discipline peu connue et pourtant de premier ordre. Piaget (1967) définit l'épistémologie en première approximation comme « *l'étude de la constitution des connaissances valables* ». Elle s'occupe de la nature de la connaissance, ses formes, ses sources et ses limites : c'est en quelque sorte la science des sciences.

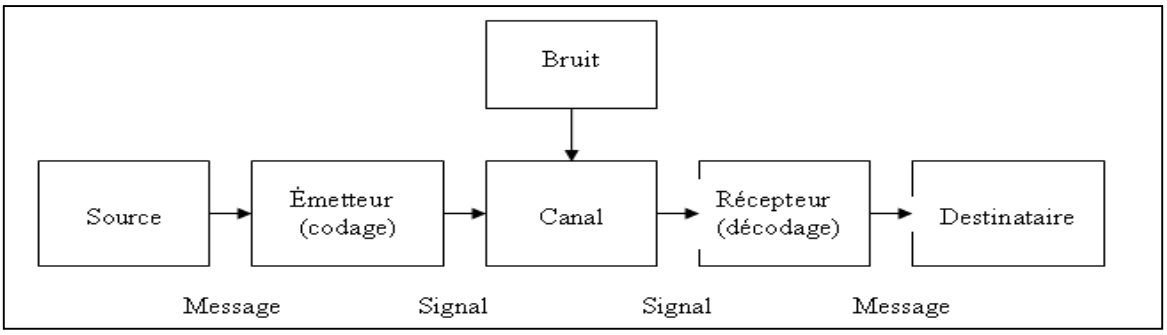


Figure 1 – Modèle de Shannon et Weaver (1949)

L'information, telle qu'elle est communément comprise, entretient un lien naturel avec l'épistémologie. L'information est un concept important lié à celui de la connaissance ; ce sont des informations que nous recherchons dans un livre ou sur Internet, que nos instruments de mesure nous renvoient, ou qui nous sont livrées encore dans le journal télévisé du soir. Cette connexion entre information et connaissance a encouragé les philosophes à utiliser les théories mathématiques de l'information pour formuler des théories de la connaissance plus précises (Dretske, 2008).

Pour cela, le modèle de Shannon et Weaver (1949) est un bon début (Figure 1). Il peut être résumé de la manière suivante : Un émetteur (ou source), grâce à un codage, envoie un message à un récepteur (ou destinataire) qui effectue le décodage dans un contexte bruité. Il est possible de faire une analogie entre « *Source* » et « *Destinataire* », d'une part, et « *Connaissances* » et « *Chercheurs* », d'autre part. En effet, c'est seulement lorsqu'un chercheur est connecté aux faits de façon appropriée −c'est-à-dire que le canal de communication entre les deux entités est fiable− que les faits peuvent être connus. Malheureusement, l'information, dans son cheminement, rencontre nécessairement du bruit, qui peut porter atteinte à sa qualité après réception. Pour Shannon, il est possible de se protéger du bruit grâce au principe de « *redondance* ». Par exemple, dans la plupart des langues, il est possible de comprendre une phase même en l'absence de certaines lettres, syllabes ou mots. Dans certains cas, le niveau de bruit ambiant est tel qu'il faut utiliser un alphabet spécial pour communiquer : c'est le cas du domaine de l'aviation où l'on utilise Alpha pour A, Bravo pour B, Charlie pour C, etc. pour faire passer un message sur un canal de piètre qualité. La théorie de Shannon appelle redondance tout ce qui dans le message apparaît en surplus. Un code redondant est ainsi transmis de manière plus fiable... mais possédera, en contrepartie, une capacité moindre (quantité d'information transmise). Il y a donc un arbitrage à faire entre la concision du codage et la probabilité de décodage correct : si on prend un codage sans aucune redondance, le décodage est sensible à toute erreur ; si le codage est très redondant, il n'est pas possible de faire passer beaucoup d'informations. Nous verrons qu'il en est de même avec la recherche : c'est en multipliant les observations d'un même phénomène qu'il est possible d'en réduire l'erreur de mesure… même si cela se fait aux dépens du nombre plus réduit de phénomènes à observer.

Le parallèle entre la théorie de l'information et la psychométrie[3] fut franchi par Lee Cronbach, l'un des psychologues le plus influent de tous les temps (Kupermintz, 2003). Nous lui devons

[3] La psychométrie est la discipline étudiant l'ensemble des techniques de mesures pratiquées en psychologie, ainsi que les techniques de validation de ces mesures

des travaux sur le concept de fiabilité, dont découle son célèbre Alpha de Cronbach (1951), l'un des indices de mesure de la cohérence interne le plus utilisé en psychologie. Il reconceptualisa également la théorie de la fiabilité dans la très impressionnante théorie de la généralisabilité (Cronbach, Rajaratnam, & Gleser, 1963) et prépara le terrain pour cinquante ans de labeur sur la notion de validité (Cronbach & Meehl, 1955). Vers la fin de la seconde guerre mondiale, Cronbach servit en tant que psychologue militaire à l'école de la Marine de San Diego où il mit au point des exercices de difficulté croissante pour aider les recrues à détecter les sous-marins. C'est à ce moment-là qu'il découvrit la théorie de l'information de Shannon. Il pensa initialement qu'elle pourrait s'appliquer à ses travaux sur la fiabilité et la validité. D'un point de vue statistique, cela s'est avéré ne pas être vrai, la théorie l'ayant plutôt conduit aux travaux de Wald sur l'utilisation de modèles statistiques dans la prise de décision (Shavelson, 2009). L'impact sur lui fut plutôt épistémologique.

Dans la théorie classique de la mesure en psychométrie, toute mesure d'un attribut peut être décomposée en deux éléments : « *le score vrai* », qui représente le score réel de l'attribut et « *l'erreur de mesure* » inhérente à la méthode de recueillement. Pour le chercheur, le score vrai représente l' « *information* » que l'on cherche à récupérer, et, l'erreur de mesure, le « *bruit* » que l'on cherche à minimiser. Shannon montre qu'un compromis entre la « *fidélité* » et la « *bande passante* » de l'information est inévitable dans tout système de communication (1949). Ainsi, dans le domaine des sciences, tout praticien ou chercheur doit choisir entre « l'étude d'une variable étroitement définie » et « *l'exploration sommaire de nombreuses entités séparées* » : c'est le dilemme fidélité-bande passante (Hogan & Roberts, 1996). Cronbach suggère que l'idéal psychométrique est une évaluation de grande fidélité et de faible bande passante (Cronbach, 1960). Il définit la notion de bande passante comme la quantité ou la complexité de l'information que l'on cherche à obtenir dans un temps donné et associe la fidélité aux termes de fiabilité, validité et d'utilité décisionnelle (1960, pp. 600–608). En suivant Cronbach, on peut placer la fidélité dans un continuum allant de « *haute* » à « *basse* » et la bande passante dans un continuum de « *large* » à « *étroite* » (Hogan & Roberts, 1996). Cronbach prend pour exemple d'évaluation hautement fidèle le test d'aptitude à l'université, car cette situation tente de répondre à une seule question avec de nombreux éléments fortement corrélés entre eux. Comme exemple de large bande passante, Cronbach cite le « *Thematic Apperception Test* » (TAT) où plus de 40 variables peuvent être évaluées à la fois (Cronbach, 1960). Chaque entreprise de recherche doit donc, selon les besoins de la recherche, se situer sur le continuum « *Fidélité-Bande passante* » de manières différentes, c'est-à-dire opter pour : « *une recherche riche en information mais peu fiable* », « *une recherche pauvre en information mais fiable* », ou « *un compromis entre les deux* ».

Ma mission au sein des Bells Labs

Soixante ans après Shannon et l'invention des transistors, vers où se sont tournées les préoccupations des Bells Labs ? Il est certain que l'ascension technologique dans le champ des systèmes d'information et de communication n'est pas prête de se terminer, elle semble même s'accélérer. En effet, le domaine réclame toujours plus de puissance de calcul, de rapidité, de mobilité, d'intégration et d'ubiquité. De ce fait, il nécessite, comme auparavant, l'appui

d'ingénieurs, de mathématiciens et d'informaticiens. Néanmoins, cela n'est plus suffisant dans toutes entreprises ayant la prétention de peser sur la scène internationale du numérique. En effet, l'informatique a quitté depuis longtemps les salles aseptisées des premiers laboratoires pour rejoindre les entreprises, puis de conquérir nos foyers et nos villes. Pour continuer à prospérer, elle a été dans l'obligation de tenir compte des caractéristiques de nouveaux acteurs. Ainsi, dans le nouveau domaine de l'Interaction Homme-Machine (IHM), trois paradigmes de recherche se sont succédés afin de permettre aux systèmes de s'adapter aux exigences de l'époque.

Si la mesure de l'information était le saint Graal dans les années 50, cela serait aujourd'hui la mesure de la qualité de l'interaction avec un système d'information, la fameuse « *User Expérience* » (UX), plus impalpable encore. En effet, à côté de la recherche fondamentale, matérielle et logicielle, on assiste à l'aggiornamento d'une recherche dite « utilisateur » dans les structures de recherche et développement des grands acteurs du numérique. Dans certaines structures, c'est elle qui pilote (« *Market Pull* »); dans d'autres, elle s'occupe de la recherche d'applications potentielles pour les nouvelles avancées technologiques (« *Technology Push* »). La recherche utilisateur aux Bells Labs se situe entre les deux. Il s'agit, d'une part, d'évaluer le potentiel des nouvelles briques technologiques issues de la recherche applicative des Bell Labs et de les instancier dans des concepts numériques innovants. D'autre part, il s'agit d'anticiper les besoins des clients pour développer les architectures clés de demain. C'est dans ce contexte, qu'une équipe, l' « *Application Studio* », a vu le jour en septembre 2010 au sein du domaine applicatif des Bells Labs. Dédiée en partie au maquettage et au prototypage rapide, elle avait pour objectif l'instanciation rapide des technologies issues de la recherche informatique afin d'en découvrir les applications potentielles. Néanmoins, il manquait encore dans l'équipe des ressources pour évaluer les créations résultantes. C'est dans ce contexte que j'ai été intégré dans l'équipe. Ma mission a été de développer des méthodes d'évaluation pour des applications de communication innovantes.

Toutefois, l'*Application Studio* ferma ses portes deux années plus tard. Cela n'a pourtant pas porté atteinte à mon travail. La question de la mesure était encore au cœur des préoccupations des chercheurs. L'équipe cherchait intuitivement à quantifier un état qui traduisait la qualité du lien entre un utilisateur et un dispositif de communication innovant. Cet état fut nommé en interne « immersion ». Le mesurer fut ma nouvelle mission. Pour mes supérieurs, cela pourrait être le nouveau Graal à poursuivre. Tout comme Shannon avait rendu possible la mesure de l'information, développer un « Immersion-mètre » (pour reprendre l'expression d'un supérieur) pourrait révolutionner le domaine. Sans en faire mon seul objectif, je peux affirmer aujourd'hui que ce terrain de recherche m'a permis de nourrir mon travail de thèse, bien que plus large.

En conclusion de cette partie introductive, je tiens à souligner que cette thèse est résolument méthodologique et cherche à faire honneur à ce champ de recherche, d'une importance souvent méconnue. C'est un sujet que j'affectionne particulièrement et, par chance, c'est une passion que je partage avec mon directeur de thèse. La prochaine partie introduira donc la nature et les enjeux de la recherche méthodologique. De plus, le contexte industriel de la thèse exige de cerner précisément les enjeux actuels du domaine des applications de communication innovantes. C'est pour cela que la partie qui suivra portera sur l'évolution qu'a connue le monde du numérique. Cela nous permettra d'ancrer le travail de la thèse sur le développement de méthodes d'évaluation adaptées à l'état actuel du domaine. Nous verrons alors que le travail

épistémologique initié par Cronbach nous sera d'un grand secours pour combler certaines lacunes laissées par une ouverture sans précédent du champ des nouvelles technologies de l'information et de la communication (NTIC).

CHAPITRE 2 : LA METHODOLOGIE ET L'EPISTEMOLOGIE

« *Une illusion de moins, c'est une vérité de plus*. »
-- *Alexandre Dumas fils* (1824 - 1895)

La réalité défie nos sens et notre raison intuitive

Nous reconnaissons tous la valeur du savoir accumulé par l'homme, autrement que pour notre seule curiosité. Après tout, c'est à partir de ces connaissances que nous avons réussi à dominer notre environnent. Mais qu'en est-il de la valeur des méthodes utilisées pour les faire jaillir? Beaucoup pensent naïvement qu'il suffit d'un bon sens de l'observation et de beaucoup de patience pour faire avancer l'état des connaissances. Mais pour cela, peut-on réellement nous fier à nos capacités innées et notre sens commun ?

Prenons l'exemple du sens le plus exploité par l'homme : la vision. Suffit-il de « voir pour connaitre » ? Notre connaissance actuelle accumulée dans le champ de la psychophysique, associée aux approches gestaltistes et constructivistes, nous montre à quel point cela serait dangereux. Elles nous dévoilent à quel point notre vision est limitée, relative, variable et biaisée. Elle est limitée car restreinte dans le spectre de la lumière visible, c'est-à-dire aux longueurs d'onde comprises entre 390 nm et 780 nm. L'au-delà, l'infrarouge, et l'en deçà, l'ultraviolet, lui sont donc inaccessibles. Les mécanismes de la vision déforment également le réel, comme le montre l'abondance des illusions visuelles existantes. Par exemple, dans l'illusion de Titchener (Figure 2), le rond central de la figure de gauche paraît plus grand que celui de la figure de droite alors qu'ils sont de taille identique (B. Roberts, Harris, & Yates, 2005). Dans l'illusion de Müller-Lyer (Figure 3), c'est la ligne avec les flèches vers l'intérieur qui parait la plus grande ; alors que dans l'illusion horizontale-verticale (appelée également illusion en T), c'est la ligne horizontale qui parait plus petite que la verticale (Figure 4). Ces quelques exemples illustrent uniquement le biais d'estimation de la taille d'objets. D'autres exemples de déformation pourraient être donnés pour la courbure des arcs de cercle, la perspective, la division de l'espace, etc. La vision est également relative. Par exemple, dans l'échiquier d'Andelson (Figure 5), la case B semble plus claire que la case A alors que les deux cases sont

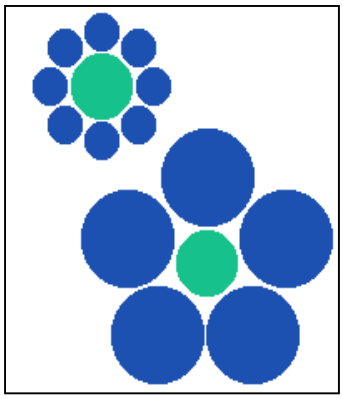


Figure 2 – Illusion de Titchener

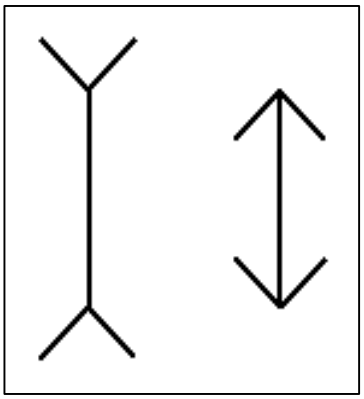


Figure 3 – Illusion de Müller-Lyer

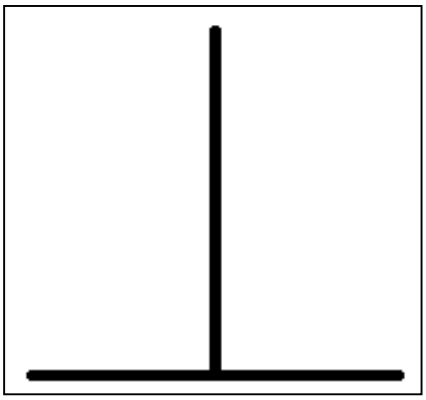


Figure 4 – Illusion horizontale-Verticale

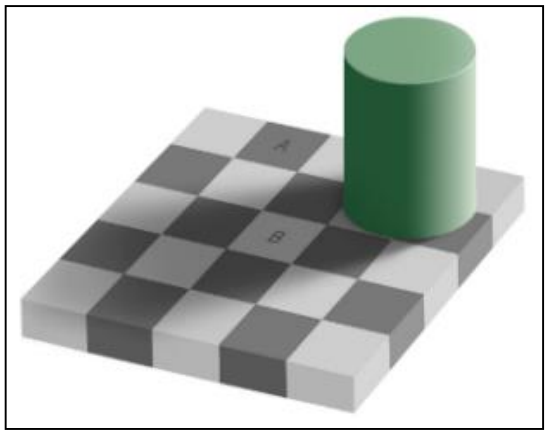


Figure 5 – Echiquier d'Andelson

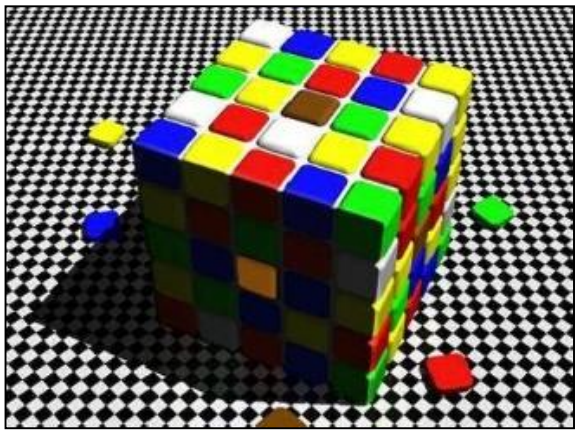


Figure 6 – Variante de l'échiquier sur la couleur

identiques ; Idem dans une variante portant sur la couleur (Figure 6), les cases centrales de chaque carré du cube semblent différentes alors qu'elles sont également de la même couleur. Ces illusions montrent que la perception d'une forme, de la taille ou d'une couleur, dépend de l'environnement et que l'œil n'accède pas à une sensation absolue du réel. La vision est également variable et plastique, car un même stimulus peut provoquer différentes réponses chez un même sujet, dépendant de son état interne (sommeil, état d'excitation, …) ou de son âge (amélioration avec la maturation des structures visuelles, perte avec la vieillesse). Elle est différente également d'un individu à l'autre, voire d'une culture à l'autre. Par exemple, les Occidentaux paraissent avoir une illusion de Müller-Lyer plus forte et une illusion du T renversé moins forte que d'autres groupes ethniques comme certaines ethnies africaines. En effet, puisque nous vivons dans un monde où il y a beaucoup de formes géométriques avec des angles droits (immeubles, murs verticaux, plafonds horizontaux, ...), nous avons une très forte tendance à surestimer les angles aigus et à sous-estimer les angles obtus : cela explique pourquoi nous sommes plus sensibles à l'illusion de Müller-Lyer. De même, pour certains peuples africains, vivant encore dans la savane ou dans le désert, l'illusion du T renversé est plus impressionnante que pour nous, car souvent le relief est plus plat et plus porté sur l'horizon. Cela réduit leurs habitudes d'estimer les lignes verticales (Segall, Campbell, & Herskovits, 1966). Ainsi, Segall (1979) déclarait que *« la perception des gens est façonnée par les inférences qu'ils ont appris à faire pour fonctionner de la manière la plus optimale qui soit dans l'environnement particulier où ils vivent. Ce que l'on peut en tirer comme conclusion est que nous apprenons à percevoir en fonction de ce que nous avons besoin de percevoir. En ce sens, l'environnement et notre culture conditionnent nos habitudes de perception* » (p. 93).

Figure 7 – Illusion d'Aristote tiré d'une gravure parue dans La Nature (1881, n°1, p.384)

Ces quelques illustrations ne portent que sur le sens de la vision. Or, ce sont tous nos sens qui nous trompent. En effet, il existe également tout un florilège d'illusions auditives, tactiles, gustatives, temporelles… voire cognitives, bien qu'il soit difficile de dire quand termine la perception et où commence la cognition (Ninio, 1998).

Comme illustration d'illusion auditive célèbre, nous pouvons citer la gamme de Shepard qui débouche sur un paradoxe, puisque d'un son à l'autre, la progression de hauteur musicale se fait toujours dans le même sens, et, pourtant, les sons se répètent à l'identique. Pour le sens du toucher, l'illusion de Thaler montre qu'une pièce de monnaie froide placée sur le front paraît plus lourde qu'une pièce chaude ; et l'illusion d'Aristote (Figure 7) –l'une des plus anciennes− souligne que, lorsque l'on roule une bille entre le majeur et l'index croisés, nous avons l'illusion de faire rouler deux billes. Cela nous montre à quel point notre perception est une construction mentale qui reflète plus ou moins fidèlement la réalité. Cet exposé est loin d'être exhaustif, et nous pourrions sans mal le prolonger par les non moins nombreux biais cognitifs, qu'ils soient attentionnels, mémoriels, motivationnels, sociaux-cognitifs, etc. Cela démontre que, même face à des situations plutôt simples et communes, le crédit que l'on peut apporter à ce que nous renvoient nos sens et notre rationalité est fragile.

Or, si nous constatons autant d'erreurs lors de situations si ordinaires, il n'est pas étonnant qu'il soit si dur de nous représenter des vérités beaucoup plus lointaines de ce que nos sens et notre pensée ont l'habitude de traiter. En effet, comment s'imaginer que la terre est ronde et tourne autour du soleil à une vitesse approximative de 30 km par seconde ? Comment penser que les atomes sont constitués essentiellement d'espace vide, et que la sensation de solidité que nous inspire le monde autour de nous n'est que pure illusion ? Comment se représenter le fait que l'espace et le temps ne sont pas absolus, mais relatifs, tant pour l'observateur que pour la chose observée, et que plus vite on se déplace, plus ces effets s'accentuent ? Nous voyons ici clairement la limite de nos sens et de notre pensée intuitive. Notre instinct nous dit que le temps est éternel, absolu, immuable – que rien ne peut troubler son écoulement régulier. Or, selon Einstein, le temps est variable et toujours changeant. Notre cerveau est dans l'incapacité de se représenter une dimension comptant trois portions d'espace et une portion de temps, entrecroisées comme les fils d'un tissu. Il en va ainsi de nombreuses théories physiques modernes que nous avons tout le mal du monde à appréhender, physiciens y compris. Bryson (2003) en donne un grand nombre d'exemples. Par exemple, sur la théorie des quanta, Bohr déclara un jour que quiconque n'était pas saisi de vertige en l'entendant parler pour la première fois, montrait simplement qu'il n'en avait pas compris le premier mot. Interrogé sur la façon dont on pourrait visualiser un atome, Heisenberg se borna à répondre : « *N'essayez pas* ». James Trefi concluait que les scientifiques avaient rencontré « *un domaine de l'Univers que notre cerveau n'est simplement pas équipé pour comprendre* ».

Certaines vérités nous sont simplement impossibles à imaginer. En effet, nos sens et notre cerveau n'ont jamais été programmés pour voir le monde tel qu'il est, mais pour assurer la survie de l'organisme qu'ils ont à leurs charges. Nos sens sont calibrés dans ce but : nous ressentons le froid et la brûlure quand nous sortons de notre zone de confort de chaleur ; et notre sens des couleurs a été initialement créé pour nous attirer vers les fruits mûrs et les feuilles tendres. Issues d'une longue évolution, nos capacités de perception et de réflexion ne s'attachent qu'à ce qui nous sert et selon une forme qui nous est utile. Sur la vision, Ramachandran (2011) exprime que son but « *n'est pas de comprendre parfaitement le monde, mais suffisamment bien et assez vite pour survivre, en laissant derrière nous le maximum de bébés* ». Du point de vue de l'évolution, c'est la seule chose qui compte. Ainsi, notre cerveau n'hésitera pas à déformer, compléter ou filtrer le réel si cela présente un quelconque avantage

de survie. Sur ce sujet, le besoin de filtrer le réel est celui qui s'explique le plus facilement : les stimulations du monde extérieur sont trop nombreuses pour toutes être traitées. De plus, nos filtres attentionnels sélectionneront en premier les informations les plus importantes pour réaliser nos buts ou les plus prioritaires, comme celles provenant de nos capteurs de douleurs. Nous éliminons également très tôt ce qui ne nous sert pas ou peu. Par exemple, nous savons que les bébés détectent les catégories phonétiques de leur langue entre six et douze mois, et cessent de distinguer les contrastes qui ne sont pas pertinents pour eux, comme le contraste entre « r » et « l » pour des bébés japonais (Werker & Tees, 1984).

Nous voyons naturellement des liens ou des formes là où il n'y a que hasard ou bruit. Ainsi, des chercheurs ont montré qu'il était parfaitement normal de voir « *le visage de Jésus dans un toast* ». Ils ont montré à une vingtaine de participants des images générées aléatoirement par ordinateur, constituées de formes indiscernables, en leur faisant croire que 50% d'entre elles contenait un visage ou une lettre de l'alphabet anglais ; Les participants virent des visages ou des lettres illusoires dans 34% et 38% des cas respectivement (Liu et al., 2014). La capacité à déceler quelque chose dans un environnement bruité est toutefois vitale dans la nature. Notre capacité à grouper et à distinguer des caractéristiques similaires a probablement évolué pour repérer des objets camouflés ou cachés dans notre environnement (Ramachandran, 2011). Elle permet de réagir au plus vite et de ne pas se faire manger par un prédateur se cachant dans un feuillage (Figure 8). Le problème, c'est la grande quantité de fausses alertes engendrées par ce processus. Dans la nature, cela n'est pas bien grave car il vaut mieux détecter la présence d'un danger alors qu'il n'y en a pas, plutôt que le contraire. Néanmoins, dans l'entreprise scientifique, qui construit son savoir sur l'existant, c'est l'inverse. Il est plus grave de voir quelque chose alors qu'il n'y a rien plutôt que de ne rien voir alors qu'il y a quelque chose.

Nombreux de ces biais viennent de mécanismes automatiques qui nous permettent de réagir vite, même avec des informations parcellaires. Par exemple dans l'illusion visuelle proposée par le physicien victorien David Brewster (Figure 9), nous partageons l'interprétation que les sphères du haut sont en relief alors que les deux autres sont creuses. Pourquoi ? Car, face à une situation perceptive ambiguë, notre cerveau tire ici une inférence sur une donnée qui nous est cachée : la source de la lumière. La source de lumière venant généralement d'en haut (le soleil), notre cerveau choisit donc l'interprétation la plus probable, c'est-à-dire que les deux sphères du haut doivent être en relief. Cette inférence d'ordre statistique fait partie des opérations élémentaires, automatiques et inconscientes de notre cerveau. Elle s'applique à toutes sortes de

Figure 8 – ***Lion caché dans le feuillage (Ramachandran, 2011)***

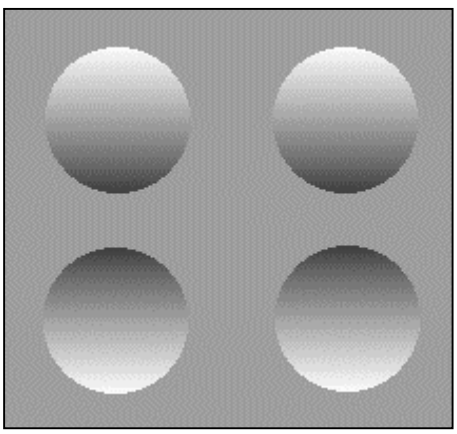


Figure 9 – ***Illusion de Brewster***

domaines de la cognition: perception, action, apprentissage du langage, reconnaissance des mots, inférences sur l'esprit des autres, etc. (Stanislas Dehaene, 2012). Confronté à des ambiguïtés, notre cerveau est capable de décider par lui-même, en se basant sur des probabilités acquises par l'expérience, et ceci sans même que nous en ayons conscience. Sur ce thème, Kanhneman (2011) avance l'hypothèse que l'esprit humain opère avec deux modes de pensée. Le système 1 fonctionne de manière automatique, rapide, et sans que l'on en ait conscience. Il ne demande aucune concentration, aucun effort: c'est le mode de fonctionnement par défaut de notre esprit. Pour cela, il utilise massivement des heuristiques, ce qui le soumet à un grand nombre de biais : surpondération des petites probabilités et sous pondération des grandes, paréidolie[4], etc. Le système 2, plus lent, moins mécanique, n'intervient que sur une base volontaire et consciente. C'est elle qui intervient lorsque l'on se concentre sur une tâche, que l'on raisonne : c'est le siège de la pensée rationnelle. Il est impossible de mobiliser ce système très longtemps, car il demande et consomme beaucoup de ressources. De plus, nous verrons plus loin que, même si l'utilisation systématique de ce système a permis de faire avancer le champ des connaissances, il ne suffit pas non plus.

La formation de l'esprit scientifique

Les illusions, nos limites de perception et de cognition ont été de tout temps exploitées dans différents domaines. Les illusions de perspective ont été exploitées dans l'architecture, les illusions de contraste de couleur dans la peinture, l'effet phi[5] dans le cinéma, et, plus récemment, l'illusion de relief, provoquée par une vibration haute fréquence, pour les interactions tactiles (Delbecq, 2014). Mais leurs compréhensions portent en elles une valeur bien plus grande. Pour Grégory (1983), les illusions doivent être acceptées « *comme des faits essentiels pour une appréhension scientifique de la nature, de la science et de l'entendement* » (p 216). Prendre en compte nos biais perceptifs et cognitifs nous permet de mettre au point des méthodes d'acquisition et de production de connaissance plus fiable. C'est là l'intérêt vital de la recherche méthodologique. Par exemple, nous savons qu'en astronomie, l'effet Purkinje[6] et l'effet Bezold-Brucke[7] peuvent fausser l'estimation de la luminosité et de la couleur d'une étoile. Connaître nos limites, c'est pouvoir y remédier en trouvant des moyens de les pallier. Il convient donc, dans un premier temps, de se détacher de nos sens naturels et d'assurer le développement d'outils de mesure plus fiables. Mais cela s'applique aussi à tous nos modes de pensées ordinaires, préscientifiques : nos constructions a priori. Il s'agit alors d'élaborer des protocoles nous mettant à l'abri de nous-mêmes, comme la passation en double aveugle.

---

[4] Illusion d'optique qui consiste à associer un stimulus visuel informe et ambigu à un élément clair et identifiable, souvent une forme humaine ou animale.

[5] Sensation visuelle de mouvement provoquée par l'apparition d'images perçues successives, susceptibles d'être raccordées par un déplacement ou une transformation

[6] L'effet Purkinje désigne la tendance pour l'œil humain de tendre vers le bleu à la fin du spectre de couleur lorsque le niveau de luminosité est faible.

[7] L'effet Bezold-Brucke désigne la tendance pour l'œil humain de confondre le rouge et le vert avec le jaune et le bleu-vert et le violet avec le bleu clair lorsque le niveau de luminosité est élevé

Mais l'entreprise scientifique doit aller parfois même encore plus loin, en allant contre les traditions, croyances et intérêts particuliers… Car il ne faut pas oublier que certaines vérités bouleversent notre vision du monde et remettent en question l'autorité de tiers. Nous connaissons l'exemple de l'héliocentrisme, développé depuis Aristarque en 280 av. J.-C ; le coup de génie de Leucippe et Démocrite qui, au Ve siècle avant l'ère commune, découvrirent l'atome sans disposer des moyens matériels de confirmer leur intuition. Nous connaissons également bien la réaction de l'ordre ecclésiastique de l'époque à de telles théories : déni, condamnation et persécution pendant des siècles. Les résistances sont nombreuses. Plus récemment, c'est la découverte de la pollution au plomb par Patterson, en 1965, qui fut escamotée par les firmes industrielles : il a fallu plus de 10 ans de combat pour que ses résultats soient acceptés. Ainsi, même avec la meilleure des démonstrations, les théories peuvent être tellement en décalage avec l'état du monde ou en contradiction avec les instances dominantes qu'il faut parfois des siècles avant qu'elles soient admises. Le scientifique n'échappe pas non plus à l'autorité et à la pression de ses pairs. À ce titre, Alexander Von Humboldt −en pensant à Agassiz qui fut raillé lors de sa présentation de sa théorie sur l'âge glaciaire− fit observer qu'il existe trois phases dans une découverte scientifique : celle où les gens nient qu'elle soit vraie ; celle où ils nient son importance ; et enfin celle où ils l'attribuent à la mauvaise personne (Bryson, 2003). Une fois, Charles Darwin déclara à moitié sérieusement que cela serait une bonne idée pour les scientifiques de mourir à 60 ans, car après cet âge, ils s'opposeront à coup sûr à toutes nouvelles doctrines (Boorstin, 1985, p. 468).

On voit ici l'importance de la preuve. Un postulat qui remet en cause beaucoup des théories bien établies doit apporter une justification d'autant plus solide : une grande théorie doit reposer sur de grandes preuves. Pour cela, la démonstration doit être sans biais, sans faille. C'est pour cela que de nombreux dilemmes philosophiques sont restés dans l'état pendant de nombreux millénaires. Ils ne commencent qu'à n'être démêlés qu'aujourd'hui, grâce à l'outillage et à la méthodologie moderne, tels qu'apportés par la psychologie expérimentale ou la neuro-imagerie. On pense à cet égard aux avancées énormes faites depuis les années 80 sur le phénomène de conscience (Dehaene, 2014) ou sur le concept associé de qualia[8], étudié ingénieusement par l'intermédiaire de la synesthésie[9] (Hubbard & Ramachandran, 2003; Ramachandran & Hirstein, 1997).

Les avancées en matière de méthodologies sont d'autant plus importantes de nos jours que nous tentons d'affiner des connaissances déjà très précises. Gaston Bachelard (1970) en décrit très précisément le mouvement : « *Jusqu'à la science contemporaine, il s'agissait de prévoir le loin en fonction du près, la sensation précise en fonction de la sensation grossière ; la pensée objective se développait quand même en contact du monde des sensations. Or, il semble bien qu'avec le vingtième siècle commence une pensée scientifique contre les sensations et qu'on doive construire une théorie de l'objectif contre l'objet. Jadis, la réflexion résistait au premier réflexe. La pensée scientifique moderne réclame qu'on résiste à la première réflexion. C'est*

[8]Phénomènes psychiques qui correspond à ce que l'on expérience lorsqu'on perçoit ou ressent quelque chose. Les qualia sont incommunicables (ex : impossibilité de décrire le rouge à un non voyant)

[9] Phénomène neurologique d'association de plusieurs sens qui ne sont naturellement pas associés (ex : forme et couleurs, textures et goûts)

donc tout l'usage du cerveau qui est mis en question. Désormais le cerveau n'est plus absolument l'instrument adéquat de la pensée scientifique, autant dire que le cerveau est l'obstacle à la pensée scientifique. Il est un obstacle en ce sens qu'il est un coordonnateur de gestes et d'appétits. Il faut penser contre le cerveau » (p. 282). L'amélioration de la qualité des méthodes de recherche permet de nous rapprocher de la vérité en nous protégeant de nos propres faiblesses. Cette quête s'est faite historiquement par une série de « bonds », en fonction des visions du monde qui jalonnaient les époques.

Les jalons et paradigmes scientifiques

Le mot science, dans l'usage qui en est fait aujourd'hui, à deux connotations : il désigne un contenu et un processus à la fois. Le contenu de la science est ce que nous connaissons grâce à elle : ce sont ses théories. Le processus de la science désigne les moyens systématiques employés pour générer ces connaissances : ce sont ses méthodes. Ce sont elles qui définissent les stratégies et procédures de recherche afin de construire et de tester les théories. Elles sont issues d'un long héritage dont la genèse résulte de toute l'histoire de la philosophie et des sciences. Le champ disciplinaire associé est celui de la méthodologie et se définit comme la science étudiant les méthodes scientifiques.

Toute méthodologie de recherche repose sur un positionnement épistémologique particulier car tout travail de création de savoir repose sur une vision du monde qui oriente le recueil ou l'analyse. Ainsi, de nombreux paradigmes épistémologiques se sont succédés au cours de l'histoire des sciences et d'autres cohabitent encore. Ils essayent de répondre à la question du « *comment nous savons que nous savons* ». Autrement dit, ce sont les théories sur la manière dont nous pouvons former les théories. Elles sont essentielles, car il n'y a « *pas de connaissance sans connaissance de la connaissance* » (Morin, 1986). De ces théories découlent des stratégies, processus et outils, permettant la production et l'évaluation de nouvelles connaissances, déclinées selon le prisme propre à chaque paradigme de recherche.

L'histoire de la construction de l'édifice scientifique est mouvementée et passionnante. La plupart de ses fondations nous semblent maintenant aller de soi alors qu'elles ont été acquises au prix des batailles terribles. En effet, « *La pensée empirique est claire, après coup, quand l'appareil des raisons a été mis au point* » (Bachelard, 1970, p. 17). On peut dire que la question épistémologique préscientifique, −c'est-à-dire s'occupant de saisir la nature et les moyens d'accéder au savoir− commence avec la philosophie antique. Dans les temps les plus reculés, la connaissance était simplement construite sur l'expérience et la pensée ordinaire ou héritée de la tradition mythologique ou religieuse. Avec le support de l'écriture, qui permet l'enrichissement du langage et de la pensée abstraite, les penseurs athéniens mettent en avant la vertu de la raison par l'usage de la dialectique, de la maîtrise parfaite du langage et de l'argumentation dès le Ve siècle avant JC. Il émerge à cette époque deux courants épistémologiques forts qui depuis lors n'ont cessé de s'affronter : le **rationalisme** et l'**empirisme**. Le rationalisme considère que toute connaissance valide provient essentiellement de l'usage de la raison. Ce courant privilègie le raisonnement a priori, qui va de l'abstrait vers le concret, comme mécanisme de production de connaissances : « *sans raison déductive point de vérité* ». L'empirisme, lui, stipule que toute connaissance provient essentiellement de

l'expérience. C'est l'accumulation d'observations qui permet d'en extraire des lois a posteriori : « *sans raisonnement inductif, point de vérité* ». Si l'équilibre entre ces courants s'est plus ou moins préservé durant l'antiquité, il se rompt durant le Moyen Âge. Le rationalisme connaît un développement fulgurant, notamment sous l'impulsion de saint Augustin (354-430). En effet, le christianisme cherchait à unir la raison avec la foi, la philosophie avec la théologie. La métaphysique chrétienne, assemblée en systèmes dogmatiques, complet, se justifiant elle-même et sans attache avec le monde externe, couplée au déni et condamnation de toutes observations du réel contredisant les Saintes Écritures, plonge l'Europe dans l'âge sombre de la rationalité en circuit fermé. D'un autre côté, la science arabe est florissante, notamment par l'impulsion des dynasties des Omeyyades (661-752) et des Abassides (750-1258) qui prolongent la science hellénistique et y incorporent des apports perses et indiens.

Il faut attendre le XVIIe siècle pour que renaissent réellement les sciences en Europe. Ce mouvement prolonge celui de la renaissance, où l'on redécouvre les textes de l'antiquité. Cela initie le mouvement des lumières, qui s'illustre par la lutte contre l'oppression religieuse et politique. À partir de ce moment, les savants associent la science à la notion de progrès et s'engagent ensemble à faire reculer l'irrationnel, l'arbitraire, l'obscurantisme et la superstition des siècles passés. On assiste ainsi à la naissance des postures épistémologiques modernes. Le rationalisme ainsi se modernise à travers l'influence énorme de René Descartes (1595-1650). Il s'inspire de la philosophie scolastique dont il partage la vocation métaphysique, mais en exclut le religieux. D'autres philosophes célèbres suivent cette voie, tels que Malebranche, Spinoza et Leibniz. Quelques-unes des idées du cartésianisme seront reprises dans la science moderne, telle que la **règle de l'évidence** et de la **méthode**. La règle de l'évidence stipule que l'on doit n'accepter comme vrai que ce qui est évident. La règle de la méthode stipule (i) une utilisation systématique de celle-ci car elle aide l'esprit à saisir les idées, (ii) l'exclusion de la religion car ses vérités, dépassant l'entendement, échappent à une méthode, et (iii) l'exclusion de la morale car elle a pour objet le bien et non le vrai.

L'empirisme se modernise également sous l'impulsion de Francis Bacon (1561-1626), précurseur de la **méthode expérimentale moderne**. Il est suivi par Hobbes, Locke, Berkeley et Hume. Son œuvre majeure de 1620, la « *Novum organum »* (nouvelle logique) a pour objet d'écarter les idoles, ou obstacles qui s'opposent aux sciences. Bacon présente quatre types de préjugés qui font obstacles à la connaissance objective : (i) les préjugés de la tribu, (ii) les préjugés de la caverne, (iii) les préjugés de la place publique et (iv) les préjugés du théâtre. (i) Les préjugés de la tribu, propre à tout homme, correspondent à notre disposition à juger les choses en fonction de nous-mêmes et non en fonction du réel. Ce manque d'objectivité s'enracine dans la nature de l'homme dont les perceptions sont à la mesure de ses désirs et non pas de l'Univers. (ii) Les préjugés de la caverne, plus individuels, correspondent à notre disposition à juger les choses en fonction de notre éducation, les événements de notre vie et de notre caractère. (iii) Les préjugés de la place publique, de loin les plus haïssables pour Bacon, procèdent des illusions et du prestige du langage. Elle découle de notre disposition à croire pour vrai tout ce que nous pouvons nommer, que cela corresponde à une entité réelle ou non. De plus, l'ambiguïté naturelle du langage crée des divergences interprétatives qui entretiennent des disputes inutiles, parfois des millénaires durant. Enfin, (iv) les préjugés du théâtre constituent la quatrième catégorie d'idole qui naît de la vénération que nous portons aux œuvres du passé

comme aux systèmes philosophiques. Deviennent vérités ce que nous voulons fortement qu'elles soient. Pour Nietzsche (1935), la logique métaphysique, cette «*méthodique falsification utile*» du réel, a toujours répondu au besoin de stabilité des hommes, de postuler un univers idéal au-delà des apparences empiriques de notre monde, matériel, changeant et chaotique. Louis de Jaucourt (1751), célèbre encyclopédiste de Diderot et d'Alembert, prolonge la pensée de Bacon sur les préjugés par l'influence des passions : « *l'entendement ne voit rien d'un œil sec & indifférent, tant l'intérêt lui en impose. Ce qui nous plaît est toujours vrai, juste, utile, solide & raisonnable. Ce qui est difficile est regardé comme inutile pour ménager la vanité, ou comme impossible pour flatter la paresse. L'impatience craint les lenteurs de l'examen; l'ambition ne peut se contenter d'une expérience modérée, ni d'un succès médiocre; l'orgueil dédaigne les détails de l'expérience, & veut franchir d'un saut l'intervalle qui sépare les vérités moyennes des vérités sommaires; le respect humain fait éviter la discussion de certaines questions problématiques; enfin l'entendement est sans cesse arrêté dans sa marche, ou troublé dans ses jugements* » (t.13, p. 284). Ainsi, le plus grand apport de l'époque dans la démarche scientifique fut celui de l'humilité : laisser une place pour le doute raisonnable, reconnaître les faiblesses de l'Homme dans sa quête de connaissance et la nécessité permanente d'être appuyé par une méthode stricte pour guider l'entreprise scientifique. Pour Bacon, la méthode est essentielle et permet de rester au plus près de la réalité : « *Ce ne sont pas des ailes qu'il faut à notre esprit, mais des semelles de plomb* », car « *on ne triomphe de la nature qu'en lui obéissant* ». La recherche des causes ultimes, obsession aristotélicienne et moyenâgeuse, est également décrite comme une entreprise stérile dans les sciences : « *La recherche des causes finales, comme une vierge consacrée à Dieu, n'enfante rien »*. Enfin, il milite pour un rapprochement de l'**empirisme** (la fourmi) avec le **rationalisme** (l'araignée) : « *Notre plus grande ressource, celle dont nous devons tout espérer, c'est l'étroite alliance de ces deux facultés: l'expérimentation et la rationnelle, union qui n'a point été formée (...) Le savant ne doit pas faire comme l'araignée, qui tire tout d'elle-même. Il ne doit pas non plus se borner à amasser des faits, comme la fourmi des provisions. Il doit grouper, classer les faits et en découvrir les lois, semblable à l'abeille qui élabore son miel.* »

Jusqu'alors, la science était liée à la philosophie, mais à partir du XVII siècle, la révolution scientifique lui permet de prendre son indépendance. Depuis lors, son influence n'a cessé de grandir en se diffusant dans un nombre de plus en plus important de domaines. Ses découvertes continueront à ébranler la place de l'Homme au-dessus de chaque chose. Cela commence avec Darwin (1809-1882) et sa théorie de l'évolution, remettant l'Homme à sa place dans le règne animal ; puis, avec la découverte de l'ADN par Watson et Crick, en 44, qui montre que toutes les espèces vivantes partagent le même code biologique ; et, enfin avec Freud (1856-1939) et sa découverte de l'inconscient, qui niera définitivement le contrôle total que l'homme pense avoir sur lui-même[10]. Face à l'efficacité phénoménale des sciences pendant ces siècles, certaines personnes succombèrent à l'idée de sa toute-puissance: c'est le scientisme. Apparaissant au XIXème siècle, cette doctrine stipule que la science a la capacité de résoudre tous les problèmes humains et doit se substituer à tout autre mode de décision. Les scientistes

[10] En effet, la psychanalyse se veut une théorie de la méconnaissance et de l'illusion puisqu'elle part du principe que chacun d'entre nous est régi par des processus dont il n'est pas l'agent conscient

veulent, selon la formule d'Ernest Renan (1890), « *organiser scientifiquement l'humanité* » (p.37). Il s'agit donc d'une confiance démesurée dans l'application des principes et méthodes de la science expérimentale dans tous les domaines. Il a atteint son apogée à la fin du XIXème siècle, puis se réduit au cours du XXe siècle, où il reste surtout vivace en URSS.

Le problème, c'est que l'arrogance d'une discipline devient en général son plus grand frein, car c'est la remise en question permanente de ces fondements qui permet sa progression. Comme le déplorait Victor Hugo, *« un fétichisme scientiste ne vaut pas mieux qu'un obscurantisme clérical* ». C'est également pour cela qu'Einstein mettra l'imagination comme valeur plus importante que le savoir cristallisé dans la science : « *L'imagination est plus importante que la connaissance. La connaissance est limitée alors que l'imagination englobe le monde entier, stimule le progrès, suscite l'évolution* »[11]. Il faut sortir des cadres de l'admis et de l'acquis pour innover. Les managers de l'innovation américains ont pour cela une expression célèbre : « *Thinking outside the box* ». Il s'agit de penser différemment, de façon non conventionnelle ou selon une perspective nouvelle. L'expression vient d'un jeu utilisé dans les années 1970/1980 par les consultants en management pour montrer à leur client qu'un cadre de pensée rigide limite la capacité à résoudre certains problèmes. La consigne du jeu était simple : relier les neuf points ci-dessous à l'aide de quatre traits et sans lever le crayon. La plupart des gens ne réussissaient pas, car ils essayaient de le faire tout en restant à l'intérieur du carré (Figure 10). Or pour réussir, il faut en sortir, « penser en dehors de la boîte », c'est-à-dire de mettre à jour les règles implicites que l'on s'est soit même fixées sans s'en rendre compte.

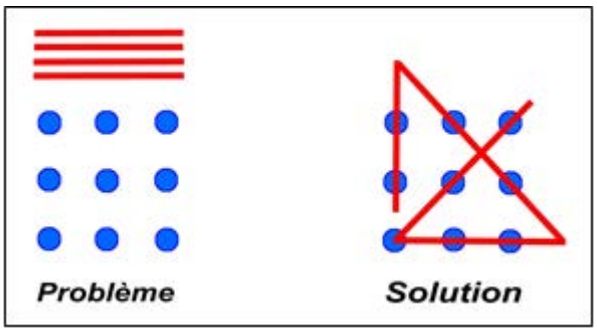


Figure 10 – *Jeu des 9 points*

L'évolution de la science ne s'est donc pas arrêtée ici. La théorie des quanta et de la relativité générale sont là pour en témoigner. L'évolution des méthodes, à un niveau épistémologique cette fois-ci, a également fait du chemin. Ce chemin, représenté par le passage du paradigme positiviste au post-positivisme, s'est nourri d'acteurs comme Popper ou Khun, qui démontrent que l'entreprise scientifique est en premier lieu un travail sur lui-même. Cela a permis de redonner l'humilité nécessaire à la science pour remettre en question ses acquis implicites et avancer. Popper (1934) est célèbre pour son critère de « ***réfutabilité*** » — en anglais, « *falsifiability* ». La démarche du savant consiste, pour Popper, non pas à prouver le bien-fondé d'une théorie, mais à essayer de la démolir, en multipliant les expériences susceptibles de démontrer sa fausseté. Ce n'est que si la théorie résiste à ces tests qu'elle peut être considérée comme scientifiquement solide … du moins jusqu'à la prochaine qui la remplacera dans le jeu des mises à l'épreuve successives. Ainsi, une proposition appartient à la science **seulement** si elle peut être réfutée par l'expérience. À l'inverse, ce qui n'est pas réfutable relèvera plutôt du domaine de la croyance ou de la mystique. L'irréfutabilité d'une théorie n'est donc pas une vertu mais un défaut (Popper, 1985). Pour Popper, la connaissance progresse par essai et élimination

[11] Citation tirée de « *What Life Means to Einstein* », George Sylvester Viereck, The Saturday Evening Post, 26 October 1929, p. 17

de l'erreur pour tendre vers la vérité, sans toutefois jamais l'atteindre ; c'est pour cela que l'on parle d'épistémologie évolutionniste (Popper, 1990). Kuhn (1962) est connu pour sa vision du progrès des sciences non pas par une accumulation linéaire, mais par révolutions, c'est-à-dire par changement de paradigme. Il s'agit d'un cycle perpétuel qui se répète tout au long de l'histoire des sciences. Pour Kuhn, une révolution scientifique intervient quand un grand nombre de scientifiques rencontrent un nombre grandissant d'anomalies, et que le paradigme de recherche en place actuellement n'arrive pas à expliquer. La discipline scientifique rentre alors en crise et de nouvelles idées sont recherchées. Ce processus va déboucher sur la mise en place d'un nouveau paradigme, une nouvelle vue commune du monde qui va remplacer la précédente. Le nouveau paradigme devra reconstruire un nouveau savoir à partir des anciennes données, car une grande partie de la connaissance d'un ancien paradigme n'est souvent plus utilisable (notion d'incommensurabilité). On notera également l'avènement d'autres paradigmes récents, tels que le pragmatisme et le constructivisme (Bachelard, 1970; Le Moigne, 1995; Morin, 1986; Jean Piaget, 1967). Ils ont apporté à l'édifice tout un appareil conceptuel supplémentaire pour comprendre les mécanismes de formation des connaissances, et qui sont autant de nuances sur la manière dont les sciences se construisent. Cette « *zététique* », art du doute cartésien, c'est-à-dire du doute sain, « *comme un moyen, non comme une fin* », ce refus de toute affirmation dogmatique, constitue le moteur du progrès d'une science normale vers l'horizon de sa scientificité : l'objectivité. Voilà pourquoi la recherche reste toujours ouverte, considérant la connaissance « *comme un processus plus que comme un état* » (Piaget, 1970), sans que jamais la science ou que ceux qui la suivent ne puissent avoir le mot de la fin, portant ainsi le dialogue vers l'infini.

Enfin, lors de l'essor des sciences humaines, le siècle dernier a vu naître d'autres épistémologies de recherche. Elles ont mis en avant certains facteurs spécifiques de recherche, affectant peu les sciences dures, mais qui sont prépondérants lorsque l'on étudie l'Homme. On a vu ainsi en sociologie s'opposer −pour schématiser simplement− deux écoles : l'école durkheimienne, « *objectiviste et holiste* » et l'école wébérienne, « *compréhensive et individualiste* ». En psychologie, c'est l'approche clinique qui s'oppose à l'approche expérimentale et comparative. On voit bien là, d'une part, l'héritage de la phénoménologie husserlienne, de l'introspection de Wundt et de l'entretien freudien sur les méthodes dites qualitatives. De l'autre côté, c'est toute l'histoire des sciences expérimentales qui forment les méthodes dites quantitatives. Nous verrons qu'il existe un paradigme de recherche qui dépasse la simple opposition de méthodes : le « *Critical Multiplism* ». Nous développerons ce paradigme dans une prochaine partie dédiée à l'évaluation multimodale et nous constaterons à quel point son utilisation sera utile dans notre entreprise. Pour conclure, nous noterons que pour la discipline traitée dans cette thèse, celle de l'IHM, nous sommes entrés dans l'ère du troisième paradigme de recherche : celui de l'expérience utilisateur. Comme Kuhn l'a très bien expliqué, chaque changement de paradigme s'effectue quand l'ancien paradigme n'arrive plus à répondre aux défis et aux enjeux du moment. De même, dernièrement, la recherche en IHM est entrée en crise. Nous sommes donc, en ce moment, dans une phase de recherche des moyens adaptés pour surmonter les nouvelles contraintes du milieu. Nous développerons dans les prochaines parties les évolutions du domaine (Chapitre 3 et 4) qui l'on poussé à développer ce nouveau paradigme de recherche (Chapitre 5). Cela nous permettra de définir les objectifs à remplir pour mettre au point de

nouvelles méthodes d'investigation du champ, tenant compte des contraintes nouvelles qu'a vu naître le domaine des IHM en ce début de siècle (Chapitre 6). Cette recherche méthodologique servira donc à fournir aux acteurs, praticiens ou chercheurs dans le nouveau paradigme des IHM, les outils nécessaires pour mener à bien leurs entreprises. Cela constituera la problématique de cette thèse. Il s'agira, bien entendu, de ne pas répondre à tous les manquements, bien trop nombreux pour une seule thèse, mais ceux que nous estimerons les plus graves.

CHAPITRE 3 : LES TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION : DU MONDE NATUREL AU NUMERIQUE

« De notre tête osseuse et neuronale, notre tête intelligente sortit. Entre nos mains, la boîte ordinateur contient et fait fonctionner ce que nous appelions jadis nos facultés: une mémoire plus puissante mille fois que la nôtre ; une imagination garnie d'icônes par millions; une raison aussi, puisque autant de logiciels peuvent résoudre cent problèmes que nous n'eussions pas résolus seuls. Notre tête est jetée devant nous, en cette boîte cognitive objectivée... Voici le savoir jeté là, objectif, collecté, collectif, connecté. »

Michel Serre

Nous pouvons dire que la révolution de l'information que nous vivons est la quatrième depuis l'apparition de l'humanité. Il y a eu avant le langage, l'écriture et l'imprimerie. Nous avons oublié que chacune de ces révolutions nous a totalement changés à chaque fois. De nouvelles capacités d'information et de communication ont fait naître de nouvelles civilisations, à chaque fois complètement différentes de celles qui précédèrent. Nous parcourrons ainsi brièvement ces différentes révolutions pour nous attarder sur la dernière et saisir en quoi nous pouvons dire que nous sommes bien en face d'une nouvelle révolution.

L'outillage naturel et la révolution du langage humain

À l'époque du stade oral, le support de l'information était le corps, avec son siège central, le cerveau. Spécialisé dans l'analyse des informations du monde extérieur, le cerveau a pour objectif principal de planifier nos actions. C'est lui qui permet notre adaptation, en tirant des règles de l'environnement et en y ajustant nos conduites. Le cerveau est donc essentiellement pour l'homme un système prédictif qui recherche les régularités dans le monde afin d'y être mieux préparé dans le futur. Le but de la connaissance peut être vu ainsi comme l'accaparation par une espèce d'une quantité de réalité pour se rendre maître de celle-ci, pour la prendre à son service et édifier un schéma de sa conduite qui suffit à nous conserver (Nietzsche, 1935). L'information favorise la préservation.

La communication, cette faculté de partager l'information, nous accompagne également depuis longtemps. Elle est d'une diversité époustouflante et se retrouve dans tout le règne du vivant. Pouvions-nous seulement imaginer, il y a seulement quelques années, que des plantes pouvaient être capables de communiquer chimiquement entre elles pour lutter contre leurs agresseurs ? [12] Les rôles de la communication sont d'une incroyable diversité et permettent une véritable interaction du vivant. Elle peut s'effectuer entre les individus d'une même espèce

[12] Voir le TED passionnant de Stefano Mancuso, neurobiologiste des plantes sur : http://www.ted.com/talks/stefano_mancuso_the_roots_of_plant_intelligence

(communication intra-spécifique) ou impliquer des individus d'espèces différentes (communication extra-spécifique). La communication intra-spécifique participe généralement à la régulation sociale et l'apprentissage. La communication extra-spécifique peut se déployer dans la symbiose entre espèces telles qu'avec les abeilles et les fleurs grâce aux couleurs et parfums. Les canaux de communication sont incroyablement variés. En effet, une communication peut être visuelle, sonore, vibratoire, chimique, cinétique, électrique, …

Nos premiers échanges communicationnels, faits de gestes et de cris, ont précédé le langage des sons articulés. Nos expressions émotionnelles sont à la base de notre premier système de communication. La communication émotionnelle, d'origine pré-langagière, est faite d'expressions faciales, posturales, et d'une modulation précise de l'intensité ou de la hauteur des sons. Elle est grandement automatique et inconsciente. Nous l'utilisons encore, mêlée au langage, pour le colorer ou l'affiner. Elle peut également être utilisée pure. En effet, un sourire ou un air menaçant se passe de mot. A un niveau fondamental et physiologique, le désir d'applaudir provient d'un débordement d'enthousiasme et d'une réaction immédiate et primitive à l'excitation, permettant la communication d'une énergie pure, non filtré par le langage et la pensée.

À partir d'un organe vocal capable de modulation (apparu pour produire des cris émus et des sons musicaux lors de parades de séduction) et préparé par la danse, le chant et les mimes, nous avons enfin formé notre premier langage. C'est un langage formé seulement de mots : un lexique sans syntaxe. Ce protolangage (Bickerton, 1990), aurait été composé de quelques mots lexicaux (verbes, noms, adjectifs) juxtaposés, sans ordre des mots bien défini, sans déclinaisons, conjugaisons, ni mots grammaticaux. Ce système de communication, télégraphique, nous aurait suffi un temps pour échanger de l'information factuelle simple.

Puis, il y a plus de 200 000 ans, est apparue notre langue humaine, celle que nous sommes les seuls à maîtriser. En effet, s'il est possible d'apprendre aux grands singes un protolangage, il nous est impossible d'aller plus loin. Notre langage, quant à lui, possède une syntaxe flexible et récursive. C'est une capacité qui ne se retrouve que chez l'Homme. C'est un langage qui permet de créer des récits complexes et de s'abroger de l'instant. Heidegger dira que l'Homme est « *l'être des lointains* » car, grâce au langage, nous pouvons nous mettre dans la peau d'autrui, en d'autres lieux et à d'autres époques, grâce au pouvoir de notre imagination. Pour certains même, tels que Hobbes ou Rousseau, le langage n'est pas simplement l'expression de la pensée ; il en est le point de départ et l'instrument. La pensée, c'est se parler, c'est « *le dialogue de l'âme avec elle-même* » (Platon). Edgar Morin va jusqu'à dire que c'est le langage qui a créé l'Homme, et non l'Homme le langage (Morin, 1973). Il précise : «*Le langage est en nous et nous sommes dans le langage. Nous faisons le langage qui nous fait. Nous sommes, dans et par le langage, ouverts par les mots, enfermés dans les mots, ouverts sur autrui (communication), fermés sur autrui (mensonge, erreur), ouverts sur les idées, enfermés dans les idées, ouverts sur le monde, fermés au monde »* (Morin, 2003, p. 31).

La révolution de l'écriture et de l'imprimerie

Une autre révolution considérable se déroula cette fois autour du premier millénaire avant J.-C. : l'écriture. L'écriture constitue le premier support d'information extérieur au corps humain,

au départ sur des pierres et des peaux de bêtes, puis sur du papyrus et du papier. Elle est d'abord signalétique, figurative, puis abstraite avec l'alphabet. Or, nous constatons qu'au moment où le couplage support/message change au cours de l'histoire, la civilisation humaine change également.

Il existe bien auparavant quelques outils pouvant amplifier ou prolonger la communication. Les divers instruments de musique ont permis d'accompagner les divers chants et processions. De plus, l'utilisation de divers moyens, tels que des signaux de fumées, tambours, miroirs réfléchissants, drapeaux, cornes ou trompettes ont été autant de dispositifs de communication primitifs ayant pour objectif de s'affranchir de la faible portée du langage humain. Ainsi, les Grecques utilisaient des relais de feu pour transmettre des messages d'invasion lors de la guerre de Trajan. Néanmoins, ces technologies de communication souffrent de deux limitations majeures : leur caractère éphémère et la pauvreté des informations transmises. Au contraire, l'invention de l'écriture, elle, aura pour effet d'immortaliser et d'enrichir les pensées.

Étonnamment, Socrate −s'exprimant par les écrits de Platon− s'attacha essentiellement à démontrer un processus d'appauvrissement enclenché par l'écriture. Il argumenta que « *cette invention, en dispensant les Hommes d'exercer leur mémoire, produira l'oubli dans l'âme de ceux qui en auront acquis la connaissance; en tant que, confiants dans l'écriture, ils chercheront au-dehors, grâce à des caractères étrangers, non point au-dedans et grâce à eux-mêmes, le moyen de se ressouvenir.* » (Platon, 1950, p. 274b–275b). Il est amusant de constater que l'avertissement des dangers à propos des technologies nouvelles ne date pas d'hier. Néanmoins, on peut voir aujourd'hui à quel point Socrate avait tort sur ce point. Le pouvoir de l'écriture d'externaliser notre mémoire sur le papier fut d'une valeur incalculable : celui de restructurer nos pensées et d'engendrer l'histoire. Une statistique permet de nous donner un indice du changement : quand une langue strictement orale contient en moyenne quelques milliers de mots, l'anglais en contient plus d'un million et s'accroît d'une dizaine de milliers par année (Gleick, 2012). Avant l'écriture, les communications sont évanescentes et locales. Les pensées sont captives de la parole, de ses hasards, de ses incertitudes, de sa dimension affective. La parole, même maîtrisée, est toujours peu rigoureuse. L'écriture, elle, est gravée dans le marbre, dure, traverse le temps, l'espace et les âges. Elle fait parler les morts et enseigne aux vivants. La persistance de l'écriture rend possible le travail aristotélicien de structuration du monde et de la pensée. En allant au-delà de la contingence, de l'instant, du concret, elle permet de construire un monde nouveau, reposant sur l'abstraction et la logique. En posant les mots, permettant aux Hommes de les réfléchir, jour après jour, l'écriture, seule, donna à la pensée cette structure systématique sans laquelle il ne pourrait y avoir de philosophie. Le langage est trop fugace pour l'analyse. La logique descend des mots écrits, depuis la Grèce, la Chine ou l'Inde, où elle fut à chaque fois développée indépendamment. Tout comme le langage a développé la pensée, l'écriture nous enseigne l'abstraction, la généralisation et la logique. Le travail du psychologue russe Aleksandr Romanovich Luria sur les populations d'illettrés en Asie centrale dans les années 30 nous en donne une démonstration (Luria, 1976). Il montra que les populations orales ne possédaient pas les catégories que même les illettrés dans les populations lettrées possèdent. Comme exemple, il leur présenta différentes formes géométriques et demanda de les nommer : pour le cercle et le carré, il obtint comme réponses « assiette, tamis, sceau, montre, lune » et « miroir, porte, maison, planche à sécher les abricots».

De plus, les illettrés n'étaient pas capables ou refusaient l'utilisation de syllogisme logique. On leur posait la question suivante : « *Dans le grand Nord, là où il y a de la neige, tous les ours sont blancs. Noyada Zembla est dans le grand Nord, et il y a toujours de la neige là-bas. De quelle couleur sont les ours ?* ». La réponse typique était : « *Je ne sais pas. J'ai déjà vu un ours noir. Je n'en ai jamais vu d'autres… Chaque lieu à ses propres animaux* ». Par opposé, un homme qui vient tout juste d'apprendre à lire réponds « selon vos dires, ils devraient être tous blancs ».

Nous ne devons néanmoins pas limiter l'analyse de l'écriture à la sphère individuelle, car c'est bien dans la société que l'on perçoit ses plus grands effets. L'invention des abstractions collectives a permis le développement du droit et de la monnaie, conduisant au développement de grandes cités et Etats. Elle a accompagné aussi le remplacement du polythéisme par le monothéisme, le développement des sciences et du commerce grâce à l'appui de textes uniques et partagés. C'est pour cela que Michel Serre va jusqu'à dire que notre civilisation est la fille de l'écriture (Serres, 2007).

La troisième révolution, dans le sillage de l'écriture, fut celle de l'imprimerie. Elle fut lancée vers 1450, quand Gutenberg eut l'idée de génie d'imprimer sur du papier à l'aide de caractères mobiles. Très rapidement, elle divisa le prix d'une bible imprimée par cinq en comparaison avec sa rivale manuscrite. Cet effondrement des prix a mis l'écriture à la disposition du grand public. Elle engendra, comme cette dernière, toute une série de bouleversements globaux : développement du commerce, évolution du droit, naissance du capitalisme, naissance de la science moderne, … Elle eut un rôle désacralisateur du pouvoir religieux, car, pour Martin Luther, « *tout homme est pape une bible à la* main ». Elle enracina une société fondée sur la liberté de pensée et le progrès technologique, malgré un dispositif de censure qui perdra rapidement de sa force en raison de l'efficacité des nouvelles méthodes de reproduction de l'information, rapides et massives.

La révolution de l'informatique

Pour l'Homme du XXIe siècle, plus aucun espace inconnu n'est à découvrir sur la Terre. Après la découverte des derniers continents, le grand défi moderne consista à relier entre eux les individus d'un bout à l'autre de la planète. Il s'agissait donc d'abolir l'espace et le temps, en imaginant une transmission instantanée de l'information ne se souciant guère des distances.

Nous savons que la distance a toujours été un problème pour transmettre des informations. Divers moyens ont été utilisés, comme les messagers coureurs de l'Empire Inca, les pigeons voyageurs ou les signaux de fumée : tout était bon pour transmettre un message le plus vite possible. Le moyen archaïque le plus intuitif a été de le confier à un congénère. Nous connaissons bien la légende du messager grec qui aurait couru de Marathon à Athènes, soit une course à pied sur route d'une distance de 42,195 km, pour annoncer la victoire contre les Perses à l'issue de la bataille de Marathon en -490 av. J.-C. Plus efficacement, et ceci de l'antiquité jusqu'au XXème siècle, l'utilisation de pigeon voyageur permit l'envoi de message court sur de grandes distances et avec une rapidité remarquable. Posséder une information avant les autres a toujours été un avantage certain. Par exemple, à l'aide de pigeons, les Rothschild

apprirent avant tout le monde la défaite de Napoléon à Waterloo. Ils s'empressent alors de racheter à la bourse des valeurs orientées à la baisse, actions qui grimpèrent de manière fulgurante une fois la nouvelle officielle. Il suffit d'une information au bon moment pour augmenter une fortune déjà considérable.

Le XIXe siècle fut celui de la « *dématérialisation* » du message. Il sera un prélude technologique d'une révolution bien plus profonde à venir. Grâce au télégraphe optique à bras (Claude Chappe, 1793), au télégraphe électrique (Samuel Morse, en 1832), et au téléphone (Alexandre Graham Bell, en 1876), les informations furent capables de se transmettre visuellement ou auditivement sans le recours à un support palpable. Les premières émissions radio, en 1920, et les émissions télévisées, dans les années 1940, complétèrent le dispositif qui permettait dès lors d'envoyer de l'information écrite, visuelle et sonore de façon quasi instantanée. Néanmoins, aucune de ces innovations ne s'apparente de près ou de loin à une révolution des moyens de communiquer, comme a pu l'être l'écriture ou l'imprimerie. Il s'agit tout au plus d'innovations technologiques, qui prolongent des moyens de communication usuels sans les transformer.

L'informatique, sortant de l'ombre durant la Deuxième Guerre mondiale, est la révolution attendue. Elle va plus loin que seulement afficher, stocker ou transférer des informations : elle les traite. Et cela sera sa principale occupation pour un temps : l'ordinateur, « *the computer* », « *compute* », c'est-à-dire calcul. En effet, pendant la Deuxième Guerre mondiale, la complexité grandissante des calculs nécessaires, comme ceux qui sont exigés par les opérations de logistique intercontinentales, dépasse de loin les compétences humaines. C'est dans ce contexte, que le premier véritable ordinateur voit le jour avec l'ENIAC, en 1946. Très imposant, l'ENIAC est constitué de plus 17 000 tubes à vide, pèse 30 tonnes et occupe une surface de 167 mètres carrés (Figure 11). Bien que de nombreux tubes brûlassent chaque jour, laissant l'ENIAC inopérant la moitié du temps, le projet fut un véritable succès. En effet, là où trois jours étaient nécessaires à un homme pour calculer la trajectoire d'une table de tir, seules trois secondes suffisaient pour l'ENIAC (Serres & Bensaude-Vincent, 1989).

Durant les vingt années suivantes, le domaine fit des progrès colossaux. Les ordinateurs passèrent des tubes à vide aux transistors, puis rapidement aux circuits imprimés, en 1961. Ces derniers furent rapidement adoptés, car ils réduisent considérablement le prix et la consommation électrique d'un système informatique. De plus, ils accélérèrent encore plus la

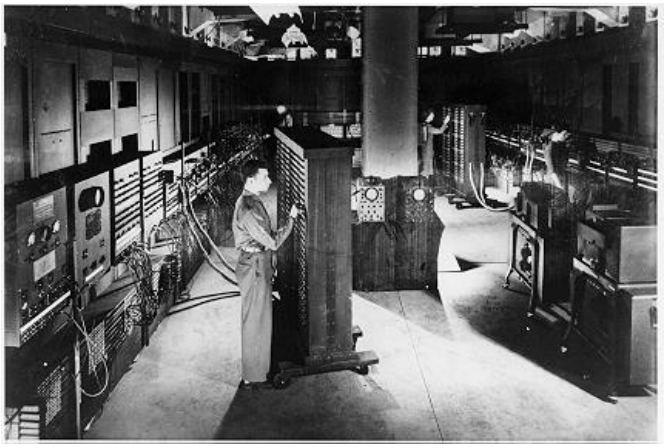


Figure 11- Ordinateur ENIAC (1946), tirée des archive d'IBM

Figure 12 – Puce PIC10 F de Microchips

course à la miniaturisation. Par exemple, Intel sorti, en 1971, une puce informatique d'une puissance égale à celle de l'ENIAC, mais avec deux différences fondamentales : un coût de 200$ pour une dimension de 12 mm² pour la puce d'Intel, contre un budget de 400 000 $ et l'encombrement d'un appartement de cinq pièces pour l'ENIAC. La micro-informatique était née. Cela était encore inimaginable deux décennies en arrière. En effet, en 1949, le magazine américain *Popular Mechanics* avait déclaré avec assurance que « *les ordinateurs du futur pourraient ne peser pas plus d'une tonne et demi* ». De même, le président d'IBM Thomas Watson avait jugé, en 1943, « qu'il y a un marché mondial de l'informatique pour peut-être cinq ordinateurs ». Comment aurait-il pu imaginer que des puces de la taille d'une tête d'épingle (figure 12) et d'une puissance bien supérieure aux monstres de l'époque pourraient être commercialisées pour moins d'un demi-dollar seulement un demi-siècle plus tard ?

Le progrès informatique continua sa course. Dans les années 80, les premiers micro-ordinateurs personnels (les « *PC* ») rencontrèrent un véritable succès, notamment avec le Macintosh d'Apple, commercialisé à moins de 2500$. À partir de ce moment, l'informatique quitta le seul domaine des entreprises pour entrer dans la sphère privée. Le reste de l'histoire nous est bien connu :

- Le développement d'Internet, du haut débit et des télécommunications mobiles ;
- L'émergence de la télématique (Rens, 1984), c'est-à-dire du mariage entre les télécommunications et l'informatique, qui permet d'acheminer indifféremment des conversations, des images vidéo et des textes écrits sous forme numérique ;
- Le développement vertigineux de l'industrie du jeu vidéo ;
- La multiplication des périphériques : PC portable, de bureau, smartphones, tablettes, …

À la fin du siècle dernier, l'informatique devint omniprésente et traite l'information sous toutes les formes. Nous sommes les premiers hommes à être capables d'accéder à tout le savoir du monde, à tout moment, et en tous lieux. Même le plus puissant des empereurs ne possédait pas ce pouvoir. Cela va même plus loin, l'informatique révolutionne radicalement l'information, la nature de nos réalités : du grec « *informa-ré* », qui forme l'esprit. En effet, la révolution informatique va au-delà de celle de l'écriture. Si l'écriture peut être décrite comme une technologie nous permettant d'externaliser notre mémoire, l'informatique externalise notre pensée et ses facultés d'acquisition, de traitement et de diffusion d'information. Ses principes sont très similaires à nos propres principes cognitifs… et pour cause : c'est sur eux que nous l'avons fondé ! En construisant un système informatique, nous nous déchargeons d'autant de nous-mêmes. Mais au contraire de Socrate, qui voyait cela comme une perte, cela nous laisse autant d'espace pour développer d'autres capacités que l'informatique ne maîtrise pas encore. Nous pouvons voir cela comme une coévolution entre l'Homme et la machine. Brangier parlera même de symbiose entre l'Homme et la technologie (Brangier, 2003). Il ajoute très justement qu'en : *« inventant des symbiotes technologiques, l'Homme crée de nouvelles ressources qui sont fondées sur ses propres qualités originelles. Il transfère, modifie et développe ses propres qualités dans la technologie. L'homme déplace dans la technologie ce qui de lui-même est programmable* » (p.8). C'est pourquoi, il poursuit en citant Simondon (1958) qui stipule que *« l'opposition entre l'humain et la technologie doit être dépassée* » (Brangier, Dufresne, & Hammes-Adelé, 2009) car *« l'anthropologie a empiriquement établi que l'anthropogénèse est empiriquement une technogenèse* » (Stiegler, 1999). En effet, comme le grand paléo

anthropologiste André Leroi-Gourhan l'a découvert il y a déjà plus de 50 ans, l'Homme commence par son extériorisation. L'Homme a toujours été celui qui s'augmente techniquement, par l'addition d'organes artificiels à ses organes naturels, dans un processus qui a commencé il y a deux millions d'années et qui se poursuit de nos jours, et c'est cela l'hominisation (Leroi-Gourhan, 1964).

En conclusion, nous pouvons dire que, cette fois-ci, nous sommes contemporains d'une révolution qui, amorcée doucement dans les années 40, ne semble pas cesser de s'accélérer. Les mêmes effets sont tant à craindre qu'à espérer que lors des précédentes révolutions sociotechniques. Les conséquences risquent d'être même plus importantes, car elles seront cette fois globales et planétaires. Certaines tendances sont déjà visibles. En effet, de nombreuses professions ne sont plus reconnaissables par leurs outillages traditionnels, car désormais l'outil de chaque métier est toujours un peu plus d'ordinateurs. Nous décrirons ces dernières tendances dans la prochaine partie et sous la forme d'un chapitre distinct, car il est important de comprendre à quel point le nombre de changements est devenu important ces dernières années.

CHAPITRE 4 : LES DERNIERES GRANDES TRANSFORMATIONS DU XXI SIECLES

《 群众是真正的英雄 - Les masses sont les véritables héros. »
Mao Tsé-Toung

Chaque société a tendance à interpréter l'histoire en fonction de son activité principale. C'est ainsi que les théoriciens du XIXe siècle découpaient l'histoire en deux parties: avant la révolution industrielle et après celle-ci. Aujourd'hui, nous assistons à une interprétation de l'histoire en fonction de l'information (Rens, 1984). Nous célébrons donc la victoire d'Hermès −le dieu des porteurs de messages et des échanges− sur Prométhée, le dieu des forgerons (Serres, 1969). Stonier (1983) cite trois changements qui caractérisent un dépassement de l'économie post-industrielle : (1) une économie majoritairement de service où prédomine une industrie de l'intelligence en lieu et place d'une industrie manufacturière ; (2) une force de travail, non pas dominée par des individus œuvrant avec des machines, mais avec des systèmes d'informations ; (3) des changements produits de manière exponentielle et non graduelle. Yann Moulier-Boutang parle ainsi de « capitalisme cognitif » (2007).

Le développement technologique et informatique, conduisant à l'âge de l'information (Zuboff, 1988), se voit donc supplanter par d'autres objectifs que ceux de l'âge précédent. L'âge industriel a consisté à augmenter la puissance physique de l'humanité sur le monde matériel, grâce à la supplantation ou l'amélioration de sa force musculaire à l'aide de machines. L'âge présent, centré sur l'information, est concerné par la puissance mentale et logique en relation aux décisions et aux comportements humains. Les machines dans l'ère de l'information, véritables « *technologies de l'esprit* » (Stiegler & Ars Industrialis, 2008), ont donc pour but de remplacer ou d'améliorer la puissance intellectuelle de l'humanité et ses capacités associées (Shackel, 2009). Cette logique se développe pour le bien −ou aux dépens− d'une nouvelle classe d'acteurs : les travailleurs de la connaissance (« *knowledge workers* »), véritable moteur de l'économie moderne (ter Heerdt & Bondarouk, 2009; Whicker & Andrews, 2004). Ce terme, introduit prospectivement par Peter F. Drucker, en 1959, désigne des employés qui « *produisent* » avec leur cerveau et vendent leur intelligence à une organisation (Drucker, 1959).

Néanmoins, cela fait maintenant plusieurs décennies que ce mouvement ne touche plus seulement les acteurs du travail. Nous sommes tous, plus ou moins, envahis par un flux d'informations qui connaît de moins en moins de frontière temporelle ou spatiale, via nos smartphones ou les nouveaux dispositifs urbains. C'est la plus grosse conséquence du développement technologique de la fin du dernier siècle, qui a progressivement fait passer l'informatique du laboratoire militaire à l'entreprise, puis du foyer à nos poches (Grudin, 2012a). Ce passage des experts aux consommateurs de masse va progressivement changer nos attentes en matière d'applications, de communication et d'information, et va progressivement envahir tous les aspects de nos vies. La consommation d'informations se densifie et se

diversifie : information sociale, « *infotainment* »[13], actualité en temps réel. La communication est possible sous toutes les formes et à un coût minime. C'est la culture de l'accès permanent, du « *quand je veux, où je veux* », et si possible gratuitement. Cette tendance s'est développée à un tel point que des polémiques apparaissent sur l'effet addictif que provoquent ces nouvelles technologies sur la consommation d'information et son rôle dans la prétendue crise de l'attention[14]. Souvenons-nous toutefois que le changement de paradigme dans l'histoire a toujours entraîné son lot d'interrogations et de résistances.

Quelles sont donc les transformations les plus marquantes du secteur à la base des bouleversements sociétaux de cette dernière décennie ? Les plus remarquables se caractérisent par un mode de gestion des applications plus communautaires et participatives (Marianna Obrist, Geerts, Brandtzæg, & Tscheligi, 2008), la convergence des canaux de communication et des contenus multimédias (Marianna Obrist et al., 2008) ; la gestion de la surcharge informationnelle et des interactions de masse ; et à la définition de produits, non plus seulement utiles et utilisables, mais visant à satisfaire une expérience utilisateur holistique (Barcenilla & Bastien, 2009).

Le Web 2.0 : l'avènement du web social et participatif

Le web a radicalement changé depuis l'avènement des médias sociaux. Les utilisateurs n'ont jamais eu autant d'occasions de s'exprimer, que ce soit par réseaux sociaux, blogs, par vidéo ou en nourrissant un savoir encyclopédique commun et gratuit comme Wikipédia. Le terme Web 2.0 a été inventé pour désigner cette nouvelle emphase (O'Reilly, 2005). Dans ce nouveau paradigme, le Web est considéré comme un outil de lecture-écriture permettant d'améliorer la collaboration et la participation des utilisateurs, afin de créer, consommer et partager de l'information (Bellucci, Malizia, Diaz, & Aedo, 2010). Bellucci et al. (2010) développent les cinq idées fondamentales derrière le Web 2.0 :

- Une architecture participative, c'est-à-dire la mise à disposition d'un ensemble de technologies et d'activités pour faciliter et promouvoir la participation, la communication et la production massive de connaissances et de savoirs.
- Un contenu généré par l'utilisateur, c'est-à-dire, produit par l'utilisateur lui-même et disponible aux autres pour y être consommé. Les nouveaux médias, numériques et connectés, en sont de parfaits véhicules. Les utilisateurs deviennent des « *prosommateurs* », c'est-à-dire producteurs et consommateurs (Rifkin, 2013). De ce fait, certains acteurs se retrouvent mis en compétition. Par exemple, le journaliste Dan Gillmor (2004) a mis en relation la prolifération des appareils photo et vidéo haute qualité et à bas prix dans les smartphones et la montée en puissance du « *journalisme citoyen* ».
- La valeur de la foule et de l'intelligence collective, basée sur l'idée suivante : *« personne n'est aussi intelligent que nous tous associés* » (Bennis & Biederman, 1998).

---

[13] Infotainment ou infodivertissement est un programme ou contenu informationnel qui inclue une part de divertissement dans l'effort d'augmenter la popularité du média avec l'audience ciblée (Demers, 2005, p. 143).

[14] Voir le fameux article Is « *Google Making Us Stoopid*? » de Nicholas G. Carr dans le magazine américain *The Atlantic* de Juillet/Aout 2008

- La « *Beta perpétuelle* » (O'Reilly, 2007), qui s'exprime par l'utilisation du Web comme une plateforme de développement et de déploiement de nouveaux services, mis en ligne rapidement et souvent mis à jour. Les utilisateurs sont ainsi mis à contribution pour l'amélioration du service, plus moins consciemment, via des sondages, des commentaires ou l'analyse des fonctions utilisées.
- L'effet de réseau, qui correspond au fait que, plus il y a de personnes qui utilisent un service, plus celui-ci prend de la valeur, et attire de ce fait encore plus d'utilisateurs (Klemperer, 2006). Dans le même sens, la célèbre loi de Metcalfe énonce que « *l'utilité d'un réseau est proportionnelle au carré du nombre de ses utilisateurs* » (Briscoe, Odlyzko, & Tilly, 2006). Le corollaire est qu'un service ou un réseau social doit atteindre ou maintenir une masse critique d'utilisateurs pour être attractif et survivre (Raban, Moldovan, & Jones, 2010).

Cette appropriation par les tout-venants de la production de contenu révolutionne les modèles économiques existants, qui s'adaptent tant bien que mal aux nouveaux usages. Joël de Rosnay (2006) va même jusqu'à opposer une nouvelle classe d'acteurs, les « *pronétaires* », aux plus traditionnels « *info-capitalistes* ». Pronétaire vient du grec « *pro* » (favorable) et de l'anglais « *net* » (réseaux). C'est un clin d'œil aux prolétaires. Pour de Rosnay, les pronétaires sont : « *une nouvelle classe d'usagers des réseaux numériques capables de produire, diffuser, vendre des contenus numériques non propriétaires* » (p.12). Véritable « *professionnel amateur* », ces acteurs s'appuient sur les principes de la « *nouvelle économie* », c'est-à-dire sur leurs capacités de créer des flux importants de visiteurs sur des sites, d'accès généralement gratuits ou à très bas prix. À l'aide d'outils quasi-professionnels leur permettant de produire des contenus numériques à haute valeur ajoutée, ces derniers rivalisent avec les « *info-capitalistes* », propriétaires des « *mass-media* » et traditionnellement seuls producteurs de contenu. Une nouvelle lutte des classes se déroule donc, sur la scène virtuelle, opposant de plus en plus d'internautes engagés et luttant contre les info-capitalistes, auxquels ils ne font plus confiance, pour s'informer, écouter de la musique, voir des vidéos ou communiquer, cela en raison des coûts trop élevés des produits et services proposés et de leur accès difficile pour les moins favorisés. La production massive et collaborative, grâce aux nouveaux outils de pouvoir des pronétaires, représente une révolution aussi importante que celle du début de l'ère industrielle. En s'appuyant sur le numérique et l'Internet, cette révolution est encore plus rapide et prend de court les pouvoirs en place dont les moyens d'action sont devenus obsolètes (de Rosnay, 2006). Le mode de communication sociétal change également (Figure 13). Il passe d'un mode pyramidal, « *top-down* », tel qu'utilisé par la télévision, la radio ou les journaux, à un mode

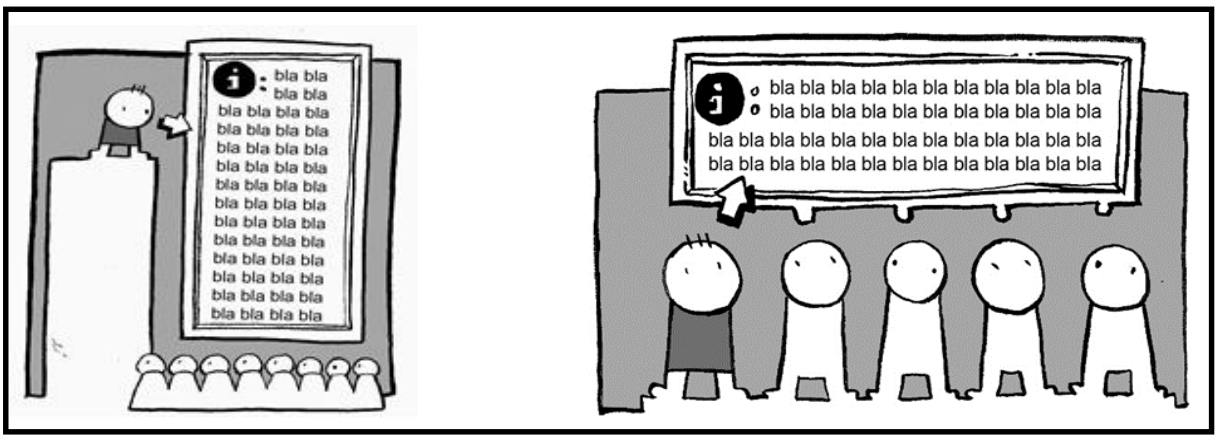


Figure 13 - D'une communication verticale à une communication horizontale (illustration tirée de diplomatie-digitale.com)

« *tous vers tous* » (« *many to many* »). Internet change donc la règle du jeu. Face aux pouvoirs publics et privés apparaît un véritable contre-pouvoir citoyen, Internet devenant le support par excellence d'information et de communication des masses. On assiste ainsi à la création de nouveaux modes d'action, tel que *change.org* qui permet de mobiliser rapidement des milliers de personnes pour une cause et d'engranger des milliers de signatures pour une pétition. De plus, les représentations digitales peuvent être dupliquées sans perte, à un coût quasi-nul et distribué à une grande échelle (Marchionini, 2008). De cette propriété unique naît une culture de la gratuité chez les « *digital natives* », et remet en cause également certains secteurs de l'économie, comme avec le *couchsurfing* ou le prêt entre particuliers, rendu possible grâce à la puissance du net. Ainsi, avec la maturation d'autres technologies telles que l'imprimante 3D, les énergies renouvelables, ou encore l'intelligence artificielle, la « *société à coût marginal zéro* » de Rifkin débouchera selon lui à une cannibalisation de l'économie capitaliste par une économie solidaire (Rifkin, 2014). Il en est de même pour Stiegler, qui stipule que l'économie actuelle devra passer du capitalisme consumériste à un nouveau modèle industriel, « *l'économie de la contribution* », qui privilégiera la passion avant l'argent (Stiegler, 2009). Le changement de modèle semble être donc un consensus chez de nombreux autres auteurs, même si ces derniers le nomment différemment : « *économie de pollinisation* » pour Moulier-Boutang (2010), « *capitalisme distribué* » pour Zuboff (2010), « *âge de la multitude* » pour Colin et Verdier (2012), *« production communautaire* » pour Benkler (2006) ou encore « *wikinomie* » pour Tapscott et Williams (2006).

La convergence numérique : la fusion des medias

Le deuxième grand bouleversement contemporain est la fusion des différents medias de communication et d'information. Il résulte directement du mouvement d'informatisation, responsable de la digitalisation de la quasi-totalité des medias et des canaux de communications actuelles. En effet, la télévision est devenue numérique, les enregistreurs vidéo ont été remplacés par leurs analogues DVD et les écouteurs cassette par des lecteurs CD ou MP3 depuis maintenant fort longtemps. Cette transition a permis la mise en place d'un mouvement plus profond : celui de la convergence de ces différents medias. Elle est permise par une autre propriété fondamentale du numérique, celle de rendre « *compatibles tous les automatismes entre eux* » (Stiegler, 2013) en leur faisant partager les mêmes protocoles, normes et langage (binaire). Cette idée de convergence, qui suggère que tous les modes d'information et de communication convergent dans un même nexus digital, n'est pas nouvelle. Le politologue Ithiel de Sola Pool identifiait déjà, à l'époque, la convergence des médias comme un processus brouillant les frontières entre les médias (de Sola Pool, 1983), notamment entre la presse, la télévision et le téléphone. Fait intéressant, les premières approches en matière de convergence ont montré que les innovations les plus fulgurantes se situaient dans les zones de chevauchement entre les différentes formes de médias ; cela semble encore être vrai aujourd'hui (de Freitas & Griffiths, 2008). Il en existe de nombreux exemples tels que :

- **la télévision digitale ou interactive** : fusion de la télévision et des technologies de l'information ;
- **les smartphones et l'avènement de la 3G** : fusion des télécommunications et des technologies de l'information ;

- **la presse en ligne ou des vidéos à la demande** : fusion de la presse et des distributeurs audiovisuels avec le Web ;
- **le serious gaming, l'e-learning et les MOOC :** fusion du domaine de l' « *Entertainmen*t », de l'éduction et de la formation avec les technologies de l'information.

L'objectif de cette convergence est de rendre disponible un plus grand nombre de produits, de contenus et de services, via une large gamme d'appareils numériques (Chakaveh & Bogen, 2007). Les mediums s'enrichissent ainsi de contenus de tous types (textuels, sonores, picturaux et cinématiques) et leurs mises en page se complexifient. En effet, le type de contenu disponible d'un périphérique n'est plus déterminé par les limites techniques de l'appareil, mais par leurs intérêts pratiques. De plus, la convergence peut se poursuivre à l'intérieur d'un même médium via l'utilisation de « *mashups* », qui permettent l'intégration de données ou de fonctionnalités provenant de différentes sources ou de services (Zang & Rosson, 2008); et via utilisation de flux RSS ou du « *Web Sémantique* », qui permettent de structurer et de favoriser l'échange de contenus informationnels. Elle est facilitée entre les médiums par la standardisation des protocoles d'échange et de représentation des données ; et par la mise au point d'outils de mise en forme automatique, permettant l'adaptation du contenu à différents appareils disposant de caractéristiques différentes en termes de résolution ou de dimension (Lin, 2006).

Ainsi, le « *Multimédia* », né dans les années 80, s'est également prolongé par d'autres approches, toujours plus intégrantes :

- **L'hypermédia :** imaginé par Nelson (1965) et instancié dans le milieu des années 90. C'est une prolongation de l'hypertexte : un média dans lequel toutes les formes d'informations sont reliées et permettent une navigation non linéaire et interactive.
- **Le Rich-media** : ou « *médias interactifs* » est utilisé dans les années 2000. Il intègre les différents médias (sons, vidéos, photos, métadonnées), et les présente de manière interactive et temporelle. La capacité du support à synchroniser l'audio et/ou la vidéo avec les autres supports est l'une des caractéristiques du « *Rich Media* ».
- **Le Transmedia** : à ne pas confondre avec le **« Cross Media »**, plus ancien et qui consiste à transposer un même contenu narratif sur différents médias (par ex. les aventures de Tintin en film, livre et dessin animé). Le transmédia, utilisé seulement depuis quelques années, consiste à décliner des contenus différents selon chaque support, contribuant au final à un programme unique. Il prend en compte les avantages et inconvénients de chaque support et forme des liens entre eux.

C'est la fertilisation croisée de produits issus de domaines autrefois bien séparés et maintenant confondus qui est à l'origine d'une multitude de services innovants. Cette tendance à digitaliser, relier et hybrider un ensemble plus grand d'objets, de personnes, de contenus et d'interfaces n'est pas prête de s'arrêter. Le développement de « *l'Internet des objets »*[15], du « *Quantified*

[15] L'internet des objets (« *Internet of Things* ») est un réseaux qui permet, via des systèmes d'identification électronique normalisés et sans fil, d'identifier et de communiquer numériquement avec des objets physiques afin

Self »[16], du « *Cloud Computing* » et du « *Big Data* » participera à ce mouvement d'agrégation. Cisco en 2013, va jusqu'à parler du développement de « *l'Internet of Everything* » (Bradley, Reberger, Dixit, & Gupta, 2013), comprenant non seulement l'Internet des objets (« *Internet of Things* ») mais également des processus, des données et des personnes via leurs smartphones et les réseaux sociaux (Figure 14). D'ailleurs, l'e-santé (liée au mouvement du *Quantified Self)* et les objets connectés (lié à l'internet des objets) font tous deux partie d'un plan de développement industriel pour la France proposé par Arnaud Montebourg le 12 septembre 2013[17].

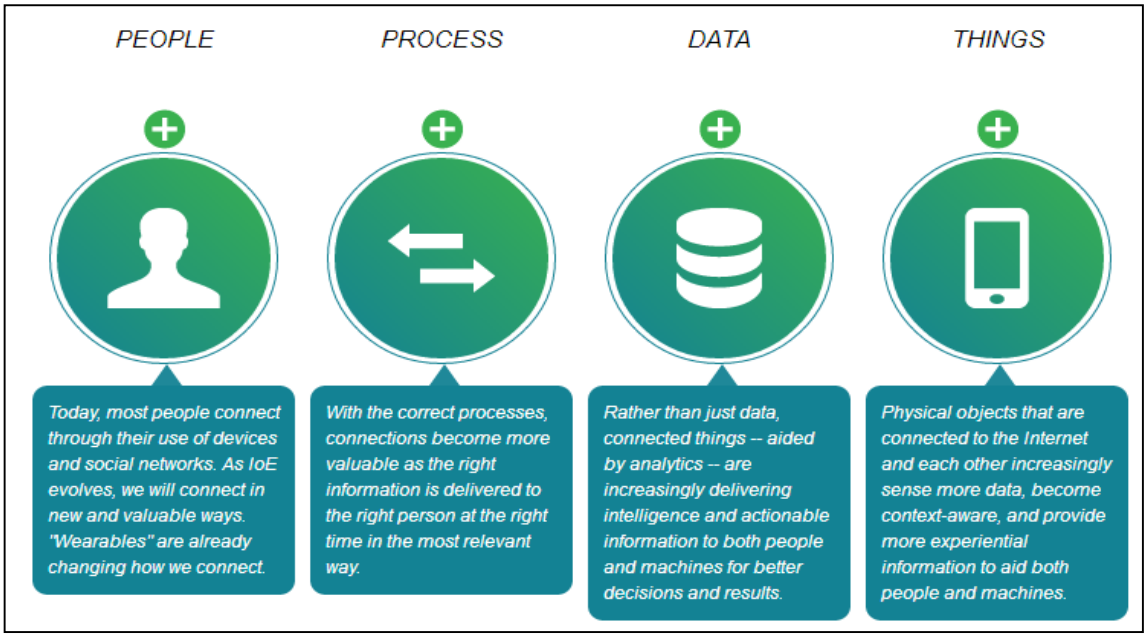


Figure 14 - Infographie Cisco (2013). Value of the Internet of Everything for Cities, States & Countries

Des entreprises aux foyers : la diversification des usages

En 1980, une ère se termine. Les ordinateurs massifs (« *mainframe* ») se voient remplacés par les mini-ordinateurs puis par la micro-informatique, entraînant un gain d'espace et d'argent. Les mini-ordinateurs permettent aux petites entreprises l'utilisation d'applications professionnelles telles que les emails, le traitement de texte et les tableurs. La micro-informatique permit l'émergence du concept d'ordinateur personnel (PC), c'est-à-dire la mise à disposition personnelle des technologies informatiques dans la sphère professionnelle ou privée. Les vagues technologiques successives, notamment avec l'avènement des interfaces graphiques et du web, accéléreront l'adoption du numérique dans la société tout entière. Avec la convergence des mediums actuels, c'est un nombre encore plus important d'activités qui se voient supporter par l'informatique. Dans le même temps, le développement fulgurant de l'industrie du jeu vidéo étendra le domaine de l'informatique dans le secteur des loisirs, des

de pouvoir mesurer et échanger des données entre les mondes physiques et virtuels (Benghozi, Bureau, Massit-Folléa, Waroquiers, & Davidson, 2009)

[16] Le « *Quantified Self* » désigne la pratique de la « mesure de soi ». Ce mouvement consiste à mesurer, à l'aide de différents capteurs, des données relatives à son corps et à ses activités pour mieux se connaître.

[17] http://www.economie.gouv.fr/files/la-nouvelle-france-industrielle.pdf

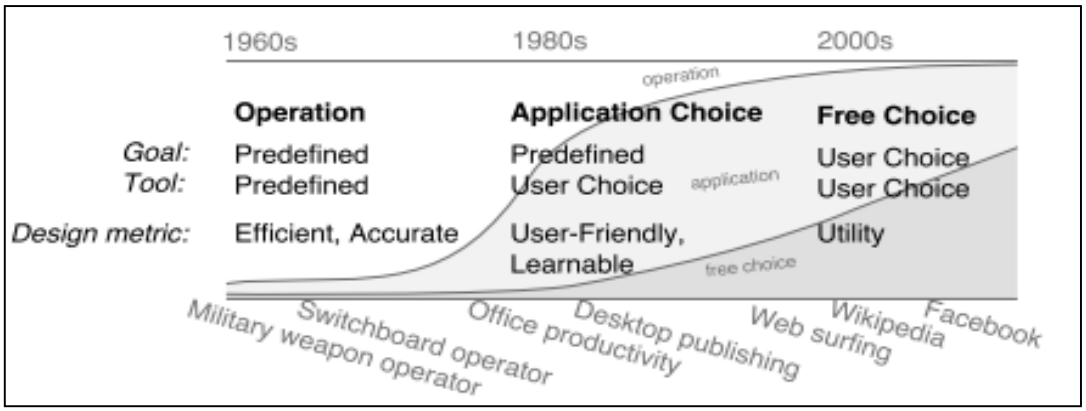


Figure 15 - Evolution du choix des utilisateurs en termes d'application (Toomim, Kriplean, Pörtner, & Landay, 2011).

plaisirs et des activités récréatives ; le commerce électronique naîtra dans les années 90 (Brangier & Bastien, 2010) ; puis les réseaux sociaux emboiteront le pas, tout comme l'e-learning, l'e-formation, l'e-santé… Ainsi, la technologie permit, non plus seulement l'assouvissement d'objectifs professionnels et pragmatiques, mais également sociaux, ludiques, esthétiques et bien plus encore. Pour chaque besoin, « *il y a une application pour cela* » selon le slogan bien connu d'Apple. De ce fait, nous sommes devenus beaucoup plus exigeants. Là où dans les années 80 nous demandions seulement à une application d'être facile à utiliser −car découlant généralement d'un contexte professionnel forcé− aujourd'hui, nous lui demandons d'être utile à l'assouvissement d'un grand nombre de nos désirs (figure 15).

Pour cela, les applications actuelles doivent répondre à un nombre grandissant de besoins ou de « *nouveaux besoins* », c'est-à-dire pas ou mal assouvis par les technologies précédentes. Un nombre important de critères rentre donc en compte pour évaluer la valeur d'une application. Ainsi, Toomim et al. (2011) propose de calculer (naïvement), l'utilité d'une application en réalisant la somme de tous les facteurs de qualité : « *Utilité = Efficience + Rapidité + Fun + Satisfaction + Récompense sociale + Beauté + Clarté + facilité d'apprentissage + Fiabilité + Intuitivité* ». Une infinité d'autres facteurs serait à rajouter dans l'équation, d'autant plus que la technologie permet d'assouvir un nombre grandissant d'activités, qui sont autant de sources motivationnelles supplémentaires.

La surcharge informationnelle et l'interaction de masse

Le mouvement de digitalisation du monde et de convergence des contenus a fait croître le nombre d'informations à notre disposition de manière exponentiel. Plus d'informations ont été produites dans les trente dernières années que depuis 5000 ans d'histoire (Craine, 2000). Certains experts estiment que les outils d'analyses actuels doivent manipuler tous les jours plus de 70 térabits d'informations, avec une augmentation annuelle de ce chiffre de 60%. Cette tendance a été accentuée par la montée du Web dans les années 1990, et plus récemment par l'avènement des réseaux sociaux et des outils du Web 2.0. Elle devrait encore s'amplifier par le mouvement de l'internet des objets, qui vise l'extension d'Internet à des choses et à des lieux dans le monde physique, grâce à l'utilisation de puces RFID. Ces avancées nous offrent l'accès à un environnement informationnel plus riche et complexe, via une plus grande variété de formats et accessible par une plus grande variété de medias et de canaux de communication.

Le risque est grand de se noyer dans cette masse d'informations. Naissent alors des situations de « *surcharge informationnelle* » (« *information overload* »), quand la quantité d'informations disponibles est supérieure à la capacité individuelle de traitement de ces informations (Bawden, 2001). Néanmoins, cette notion est loin d'être neuve, et fut déjà caractérisée explicitement comme un épineux problème durant la « *Royal Society Scientific Information Conference »* de 1948. Elle se traduisait à l'époque par une peur des scientifiques de se voir dépasser par la masse croissante de publications à suivre et déversées par la presse mondiale (Bawden, 2001). Dans les années 90, la surcharge d'informations a commencé à être considérée également comme un problème majeur dans le monde des affaires, du commerce, des milieux universitaires et de la santé (Bawden & Robinson, 2008). Actuellement, ce sont les nouvelles technologies de l'information et de la communication qui amplifient l'effet de surcharge cognitive, notamment par les systèmes de type « push », c'est-à-dire fournissant activement des informations à l'utilisateur sans aucune demande de sa part. Elles s'instancient sous la forme de notifications, d'emails, d'abonnement à des flux RSS, de chaînes ou de listes qui peuvent dépasser notre capacité à les parcourir. La nature des outils web 2.0 favorise également un paysage de l'information basé sur la nouveauté superficielle. Parce que les outils permettent et encouragent les mises à jour rapides et l'affichage de nouvelles informations, il se dégage une attente de nouveauté constante. De plus, les médias mobiles ont favorisé les *« Fast News* », informations instantanées et omniprésentes, qui sont devenues la normes pour occuper son temps libre, se sentir libre ou maintenir le lien social (Express Roularta Services, 2011). Le flux incessant d'informations et la multiplication des canaux utilisés généralisent le multi-tasking (Figure 16). Dans certains cas, le surplus d'informations d'un seul système peut faire déborder notre charge cognitive (Otondo, Vanscotter, Allen, & Palvia, 2007).

Dans ce contexte, la meilleure des armes est d'abord de ne pas se laisser submerger, en mettant en place une hygiène personnelle d'acquisition, de traitement et de stockage personnel des informations (Jones, 2007). Pour cela, de nombreux outils ont vu le jour sous la dénomination de systèmes personnels de management d'informations (« *Personnal Information Management Systems* », ou « *PIMS* » ; Boardman, 2004). Le « *Satisficing* » (seuil de satisfaction), est souvent avancé comme un moyen heuristique efficace pour faire face à la situation de débordement d'informations : cela consiste à prendre juste assez d'informations pour répondre à un besoin, plutôt que d'être submergé par toute l'information disponible (Schwatrz, 2004). Ce terme est un mot-valise formé des mots « *satisfying* » (satisfaisant) et « *sufficing* » (suffisant). L'économiste Herbert Simon y fait référence sous la notion de « *rationalité limitée* » (Simon, 1955). Les acteurs du Web ont également développé une série d'outils et de fonctionnalités afin de supporter l'interaction et l'exploration de ce gigantesque espace d'informations par les utilisateurs tels que :

- le classement, la priorisation et le filtrage d'objets ou de résultat de recherches ;
- la recommandation personnalisée (ex : eBay) ou par les liens sociaux (ex : facebook) ;
- la catégorisation des données par un processus collaboratif de tagging ;
- une meilleure présentation des données à l'aide des méta-données et de connaissances issues de l'architecture d'informations (Wurman, 1996) ;
- le développement de la « *Visualisation d'Information* » (« *Information Vizualisation* »), qui permet la représentation et la fouille de grands ensembles de données (Card, 2009);

- la mise au point de moteur de recherche ouvert permettant la sérendipité, c'est-à-dire « *la découverte de ce que l'on ne cherchait pas* » (Sandri, 2013) ;
- la mise à disposition d'informations de qualité et de synthèse (« *Slow News* »), sous une forme adaptée et captivante, grâce à l' « *Info-telling* » ; (Express Roularta Services, 2011).

En parallèle de la croissance exponentielle de données sur internet, le Web 2.0 a contribué à l'explosion de la participation des utilisateurs. Cela est particulièrement visible quand on examine les géants du Web comme Facebook, Youtube et Wikipedia. Facebook a été créé en Février 2004 comme un site de réseautage pour l'université d'Harvard et a été ouvert au grand public seulement en 2006. En décembre 2006, Facebook comptait déjà plus de 12 millions d'utilisateurs. En décembre 2009, ce chiffre passe à 350 millions (Facebook, 2009) et atteint 665 millions en mai 2013 (Facebook, 2013). C'est le deuxième site internet visité au monde juste derrière Google (Alexa Internet Inc., 2013). Après lui se situe YouTube, qui connait une progression toute aussi fulgurante. Combien d'heures de (nouvelles) vidéo sont uploadées chaque minute ? 6h en 2007, puis 24h en 2010, et 60h en 2012. Cela fait donc une heure de vidéo uploadée par seconde (Youtube, 2012). Wikipédia, en sixième position, suit un parcours similaire. Depuis sa création en 2001, le site a connu une croissance exponentielle. Il compte aujourd'hui près de 2,5 millions de pages et plus de six millions de contributeurs enregistrés, rien que dans la partie anglaise (Alexa Internet Inc., 2013). Des études ont montré que son contenu est aujourd'hui de qualité comparable à celles des encyclopédies traditionnelles (Giles, 2005), et que le vandalisme et les inexactitudes sont souvent corrigées en l'espace de quelques minutes (Kittur, Suh, Chi, & Pendleton, 2007; Priedhorsky et al., 2007). Pour que cette « *sagesse de la foule* » s'opère, un système de coordination et de communication entre les utilisateurs a dû être mis en place (Kittur & Kraut, 2008). En effet, pour que cette foule soit « sage », Surowiecki (2005) montre qu'il faut que quatre éléments soit réunis : la diversité d'opinion, l'indépendance, la décentralisation et l'agrégation. Au contraire, la sagesse de la foule n'apparait pas quand la prise de décision est trop centralisée, trop divisée ou imitative. Le processus de répartition des tâches entre un grand nombre d'individus a donné naissance au terme de « *crownsourcing* » (Howe, 2006). L'utilisation massive de la puissance de traitement humaine au service d'un système informatique, a donné lieu au concept d' « *Human Based Computing* ». Ainsi, le « *Calcul Humain* » (« *Human Computation* ») est défini par von Ahn (2005) comme « *...un paradigme utilisant la puissance de traitement humaine dans la résolution de problème que les machines ne peuvent pas encore résoudre* ». Cette nouvelle approche est particulièrement adaptée pour la résolution des problèmes complexes d'analyse de données impliquant un grand nombre de

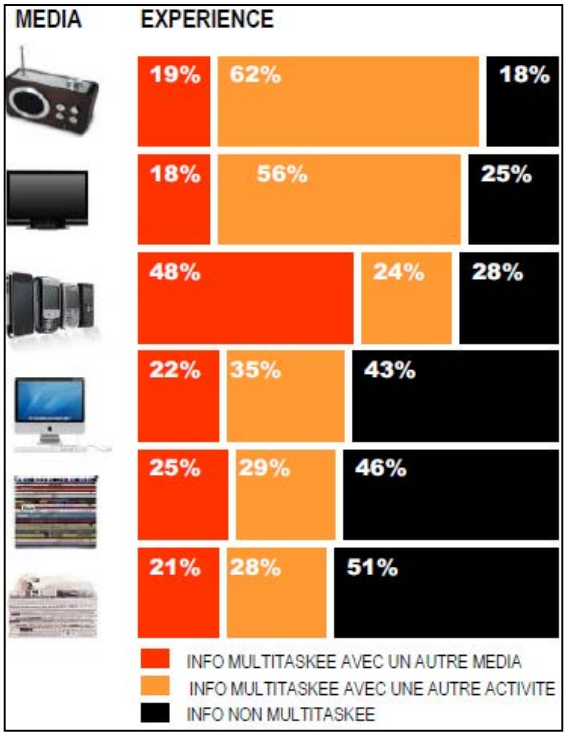


Figure 16- Généralisation du multi-tasking (Express Roularta Services, 2011)

personnes et de gros volumes d'informations. Cette force de travail collaborative a été mise au service de nombreux projets tels que :

- le classement de cratères sur la surface de Mars par des bénévoles, résultant en un travail pratiquement indiscernable de celles des géologues experts (Kanefsky, Barlow, & Gulick, 2001) ;
- l'étiquetage de contenu d'images pour en décrire certains éléments (Luis von Ahn & Dabbish, 2004) ;
- le repliement des protéines (Cooper et al., 2010), qui a permis au bout de trois semaines seulement de trouver la structure tridimensionnelle de la protéase rétrovirale du virus M-PMV, (ce qui avait bloqué les scientifiques pendant plus de 10 ans) ;
- L'analyse du contenu des notes de frais des membres du Parlement anglais par des lecteurs de la revue Guardian (Arthur, 2009).

Certains auteurs vont même jusqu'à imaginer des dispositifs logiciels utilisant la sagesse des foules en temps réel, et permettre par exemple lors de la prise d'un segment vidéo de sélectionner le meilleur cliché dans celui-ci (Bernstein, Brandt, Miller, & Karger, 2011). Les personnes sont motivées à participer à la réalisation de ces tâches en jouant sur différents leviers sociaux (célébrité), moraux (faire avancer la science ou retrouver une personne disparue) ou ludique (Quinn & Bederson, 2011). Dans certains cas, les travaux sont rémunérés à la tâche, tels que sur Amazon Mechanical Turk (mturk.com) ou microworkers.com. On parlera alors de « *cloud labour* », « *e-labour* » ou d' « *eLancing marketplaces* », ce qui correspond dans les faits à une forme numérique du travail freelance (Aguinis & Lawal, 2012).

Conclusion : L'IHM d'aujourd'hui et des Bells Labs

Le champ de la recherche en interaction homme-machine regarde plus que jamais en avant en raison du rythme extraordinaire des progrès technologiques et les attentes que cela engendre. Nous savons que nous serons en mesure de faire demain ce que nous avons à peine rêvé hier. Il est intéressant de voir à quel point les IHM ont changé au cours de ces dernières années. Cela est d'autant plus évident si l'on analyse chacun des termes du champ de l' « *Interaction Homme Machine* ».

Le « M » ou « Machine » d'« IHM » est sans conteste celui qui a le plus changé. Il a pris une multitude de formes et s'intègre dans de nombreux objets de notre quotidien. Il a su devenir tout petit et nous accompagne en tous lieux et toutes heures ; ou au contraire devenir tentaculaire en reliant des milliards de personnes en même temps. En effet, il a su s'intégrer dans l'environnement et rend notre voiture, notre maison, notre ville intelligente. Le « H » ou « Homme » d'« IHM » a vu sa portée évoluer. Il est devenu plus qu'un simple utilisateur car s'ajoute à lui le rôle de consommateurs, créateurs ou joueurs en fonction des opportunités ou de ses désirs. Leurs études ne se limitent plus d'ailleurs au seul utilisateur isolé mais également aux groupes d'utilisateurs, voire même aux foules numériques. Elle ne le confine pas non plus au monde du travail et les considère dorénavant comme des acteurs numériques quel que soit le moment de la vie. Il en résulte que le « I » ou « Interaction » d'« IHM » s'est également transformée. Elle a su s'adapter au contexte : efficace en situation de performance et plus

naturelle dans les situations communes. La technologie nous connaît mieux et nous interagissons avec elle avec moins de contrainte. Elle devient un prolongement de nous-même, et ambitionne bientôt d'anticiper nos besoins.

Les changements de nom du journal de l' « *Association Francophone d'Interaction Homme-Machine* » (AFIHM) illustre également cette évolution. Crée en 1998 sous le nom de « *Revue d'Interaction Homme-Machine* », il devient en 2008 la « *Revue des Interactions Humaines Médiatisées* » puis le « *Journal d'Interaction Personne-Système* » en 2011.

Le champ de recherche applicatif des Bells Labs puise dans les thèmes modernes du domaine. Au moment de mon intégration en 2010, les départements de recherche travaillaient sur des problématiques très diverses, comme les réseaux sociaux (de l'entreprise, du livre, …), l'immersion à distance, les mashups de communication, la vidéo augmentée ou encore l'internet des objets. De plus, l'*Application Studio*, mon équipe d'appartenance, a été chargée de produire des prototypes d'applications innovantes en suivant trois grands axes de recherche :

- ***One Million Conversations*** : cet axe a pour objectif la création d'outils de gestion des connaissances personnelles (*PKM : Personal Knowledge Management*) permettant d'extraire du savoir des conversations et des échanges professionnels en général. Il s'agit de comprendre comment représenter et manipuler ce savoir afin qu'il ne participe pas à la surcharge informationnelle qu'il cherche précisément à résoudre.
- ***Video Applications*** : cet axe a pour objectif la création de démonstrateurs techniques proposant de nouvelles expérience utilisateur dans le cadre de la vidéoconférence à distance. Cet axe vise à mettre en avant les technologies avancées d'analyse d'images (reconnaissance faciale et de mouvement…) afin de rendre plus attrayante, pour un auditeur connecté sur son ordinateur de bureau, une présentation de documents effectuée au moyen d'une vidéoconférence.
- ***Environment As A Service*** : cet axe a pour objectif la création de maquettes permettant d'appréhender de nouveaux usages en matière d'internet des objets et d'informatique « *pervasive* ». Les évolutions techniques en matière d'objets communicants permettent aujourd'hui d'envisager l'apparition de nouvelles applications non WIMP, mais le frein principal est que l'expérience utilisateur ainsi créé reste très spéculative. Cet axe vise donc à mesurer l'attrait des utilisateurs pour de nouvelles formes d'accès aux services n'utilisant pas l'ordinateur ou le téléphone, et de définir les conditions nécessaires pour que ces services soient perçus, compris et acceptés (Schmitt, Kindsm, & Herczeg, 2010).

Ces trois axes de recherche sont autant de défis pour l'évaluation d'applications, dans des domaines aussi variés que la représentation visuelle de connaissances, la réalité augmentée ou encore l'informatique pervasive et contextuelle (« *Context aware computing »*). Pour cela, il nous faut approfondir nos connaissances méthodologiques en ergonomie des IHM. La définition de ce champ académique, son évolution au cours des différents paradigmes de recherche, ainsi que les enjeux à résoudre actuellement seront développés dans la prochaine partie.

CHAPITRE 5 : LA METHODOLOGIE EN ERGONOMIE DES IHM

« *If we want users to like our software, we should design it to behave like a likeable person* »

Alan Cooper

Les méthodes et l'évaluation

Dans les sciences, les méthodes de recherche se définissent comme des stratégies particulières que les chercheurs utilisent pour collecter les preuves nécessaires à la construction et à l'évaluation des théories (Frey, Botan, Friedman, & Kreps, 1992). Elles connotent un ensemble de règles et de procédures utilisées pour la construction de toutes formes de connaissance (Daly, 2003). Chaque méthodologie se base sur une épistémologie particulière dont elle tire ses principes. Ces dernières peuvent se décomposer en éléments plus petits qui constituent un état de l'art des pratiques valides d'utilisation de la méthodologie (procédures, prise en compte de biais, contrôles, principes statistiques, …).

Nous avons vu que, tout au long de l'histoire de la science, les changements de paradigme ont grandement contribué au développement des connaissances humaines. Ils ont permis, par différentes visions du monde et d'accès à la réalité, d'opérer des sauts qualitatifs à l'intérieur d'un champ disciplinaire donné. De même, le développement des méthodologies de recherche au sein de ces paradigmes est également une source de progrès. Elle peut être linéaire, quand il s'agit d'améliorer les méthodes existantes ; ou novatrice, quand il s'agit de remplacer les vieilles méthodes lors d'un changement de paradigme. Il est d'ailleurs intéressant d'examiner le lien entre les prix Nobel décernés et les progrès méthodologiques. En 2012, Greenwald analysa les prix Nobel de physique, de chimie et de médecine décernés entre 1991 et 2011. Il montra que sur cette période de 21 ans, 82% des prix Nobel ont été décernés pour une contribution méthodologique (63 sur 77) contre 18% pour une contribution théorique (14 sur 77). Il montra également que les prix Nobel décernés pour des avances théoriques sont peu fréquents quand ils se basent sur d'anciennes méthodologies ; et, quand bien même cela est le cas, les méthodes sont souvent poussées en dehors des limites d'utilisation usuelles (Greenwald, 2012). Il avança ainsi que beaucoup de grands dilemmes en psychologie pourraient être résolus par une avancée des méthodologies de recherche. Il confirma le fait que la recherche méthodologique est un enjeu vital pour la progression des sciences. Cela est également le cas en ergonomie, d'autant plus lors d'un changement de paradigme, comme celui que l'on connait actuellement.

En IHM, l'étude des méthodes d'évaluation a un intérêt important, car ce sont elles qui permettent de juger avec validité la valeur d'une application et donc de son futur succès. Le domaine d'étude des IHM se situe à l'intersection de deux disciplines : entre celle de la psychologie et des sciences sociales, d'une part, et de celle de l'informatique et des technologies, d'autre part (Carroll, 1997). Les méthodes d'évaluation en IHM permettent d'estimer la qualité d'interactions entre ces deux entités, c'est-à-dire entre l'homme et la

machine. Il reste à définir les facteurs contribuant à la satisfaction (le modèle qualité) et la procédure pour mettre en œuvre le jugement (la méthode). Dans toute évaluation, il y a un jugement de valeur, c'est-à-dire qu'il faut déterminer ce qui est important à estimer. Cela peut être la performance d'une application, sa sûreté d'utilisation, son esthétisme, … Ce modèle peut être implicite ou explicite et peut varier en fonction de la conjoncture ou du paradigme de recherche dominant.

« *Procédure d'évaluation* » et « *évaluation* » se confondent souvent dans la littérature. Il suffit en général de voir comment les gens définissent l'évaluation pour le comprendre. Ce travail de définition est toujours intéressant, car il n'est jamais vraiment fructueux. En effet, les auteurs ne s'entendent que rarement sur les mots et les étudier nous permet en général de cerner les différentes écoles de pensée existantes. Le constat est encore plus cinglant pour Glass et Ellett (1980) : « *l'évaluation, plus que n'importe quelle science, est ce que les gens disent que c'est ; et en ce moment, les gens disent que c'est plein de choses différentes.* » (p.1). De ce fait, la définition de l'évaluation qu'ils proposent, à vouloir être très consensuelle, est assez pauvre : « *L'évaluation est un ensemble d'activités pratiques et théoriques à l'intérieur d'un paradigme largement accepté*». Ils relèvent toutefois sept conceptions intéressantes de l'évaluation, qui chacune détermine ce qui convient d'évaluer et la manière de s'y prendre:

- **(i) L'évaluation comme science appliquée** : L'évaluation dans ce paradigme est vue essentiellement comme une science appliquée. Les construits sont opérationnalisés et mesurés quantitativement. Peu de choses peuvent être appréhendées à la fois et des contrôles doivent être mis en place. L'évaluateur est vu comme un expérimentateur recherchant les causes ;
- **(ii) L'évaluation comme gestion d'un système :** l'évaluation revendique ici d'élargir le champ de l'enquête d'autant que les partisans des sciences appliquées (i) le réduisent. Ils voient l'évaluation comme l'étude d'un système complexe qu'il faut administrer à tous les niveaux. Le mode de décision est rationnel au sens wébérien du terme, c'est-à-dire qu'il doit servir les buts du système ;
- **(iii) L'évaluation comme théorie de la décision** : l'évaluation est vue comme une prise de décision en présence d'incertitudes. Elle se base sur des modèles décisionnels théoriques et statistiques ;
- **(iv) L'évaluation comme mesure de la progression vers un but** : l'évaluation est vue comme une spécification de buts et de mesure des progrès effectués pour les atteindre ;
- **(v) L'évaluation comme jurisprudence** : l'évaluation est vue comme le résultat d'un verdict s'appuyant sur un débat contradictoire, argumenté et appuyé par des preuves ;
- **(vi) L'évaluation comme description ou portrait** : l'évaluation est vue comme la description totale et ethnographique d'un système. C'est une méthode d'évaluation qui s'utilise sur le terrain, au contact des besoins des acteurs et en réponse à un environnement toujours complexe ;
- **(vii) Empirisme rationnel :** l'évaluation est vue comme un compromis entre les objectifs fondamentaux d'une évaluation et les possibilités données par la situation. De ce fait, l'évaluation est avant tout la résolution d'un problème méta-évaluatif.

Glass et Ellett (1980) montrent que la majorité des conceptions de l'évaluation sont assez dogmatiques en défendant des positions épistémologiques parfois très tranchées. La seule

exception est la position de l'empirisme rationnel, représenté par Scriven (1974). C'est cette dernière approche que Glass et Ellett défendent, car elle fait preuve de ce que les autres manquent : de pragmatisme. D'ailleurs, la définition de Scriven (1967, 1991), l'une des premières sur l'évaluation, est encore celle utilisée couramment de nos jours : « *L'évaluation se réfère à la procédure de détermination de la valeur de quelque chose, ou à la production de ce procédé. Les termes utilisés pour se référer à ce processus ou à l'une de ses parties incluent : évaluer, analyser, estimer, critiquer, examiner, noter, inspecter, juger, classer, étudier, tester, … La procédure d'évaluation implique en général une identification des standards de valeur pertinents ; quelques investigations des performances de l'objet d'étude en fonction de ces standards; et la synthèse ou l'intégration de ces résultats pour parvenir à une évaluation globale ou d'un ensemble d'évaluations conjointes* » (Scriven, 1991, p. 139).

Il faut souligner toutefois que le choix d'un paradigme d'évaluation se fonde avant tout sur les besoins d'évaluation. En effet, les méthodes d'évaluation sont des procédures utilisées en vue d'accomplir un but spécifique et donc l'approche déterminera comment les données seront récoltées, analysées et présentées (Pedro Antunes, Herskovic, Ochoa, & Pino, 2012). Par exemple, en fonction du niveau de développement d'un projet, l'évaluation devra répondre à des besoins différents. Dans cette perspective, Preskill et Russ-Eft (2005) discerne trois types d'évaluations :

- **L'évaluation de développement** : l'évaluateur est placé comme faisant partie de l'équipe de conception. Il collecte des informations pour cerner les besoins et fournis un feedback informel à l'équipe de conception afin de les aider à perfectionner le système avant qu'il soit conçu. Elle est en général effectuée en début de projet ;
- **L'évaluation formative** : son objectif est l'amélioration ou l'affinage d'un projet. Elle est généralement effectuée en cours de projet ;
- **L'évaluation sommative** : son but est de déterminer la valeur de l'objet et de permettre ainsi de prononcer un jugement évaluatif. Elle est généralement effectuée à la fin de la conception d'un projet.

Les types d'évaluations peuvent également différer en fonction de ce qui importe de mesurer (la performance pure, l'adéquation avec les besoins utilisateur, …) et des diverses contraintes qui pèsent sur le projet (budget, temps, expert disponible, …). Par exemple, le choix d'une évaluation empirique (reposant sur l'observation ou le recueil d'opinion) ou analytique (reposant sur un corpus de connaissances existantes) peut être contraint par l'expertise à disposition ou par la difficulté d'accès à la population cible.

Il est toutefois possible de distinguer deux modes historiques d'évaluation qui se sont opposés tout au long de l'histoire des sciences humaines : l'évaluation quantitative et qualitative (Tableau 1). Cet affrontement prolonge toutefois des positions épistémologiques antiques, remontant bien avant la naissance de la science elle-même: celle de l'holisme (qui se prolonge dans les postures modernes qualitatives) et celle du réductionnisme (soutenue par les chercheurs quantitatifs). La position réductionniste prit un essor considérable au moment du développement des sciences naturelles et expérimentales. Lors de l'avènement des sciences sociales, la complexité des phénomènes en jeu a provoqué la réintégration du paradigme holiste sur la scène épistémologique. L'holisme est un système de pensée qui stipule que les

caractéristiques d'un objet d'étude ne peuvent être connues que lorsqu'on le considère et qu'on l'appréhende dans son ensemble, et non pas par l'étude individuelle de chacune de ses parties séparément. À l'inverse, le réductionnisme stipule que pour connaître un objet, il faut le décomposer pour étudier chacun de ses éléments indépendamment ; puis, ayant compris leurs fonctionnements distinctement, de les réassembler afin de comprendre le système de base étudié dans son ensemble. Toutefois, même si ces positions épistémologiques s'opposent, toutes deux ont leur mot à dire dans un cycle de recherche équilibré, car elles présentent des qualités complémentaires (Daly, 2003). Cela est d'autant plus vrai dans le champ complexe des IHM. Nous verrons en effet que les évaluations en début de projet sont généralement qualitatives alors que celles qui sont réalisées à la fin sont souvent quantitatives.

Tableau 1 – La controverse quantitatif & Qualitatif sur le mode d'accès à la connaissance *(Cook & Campbell, 1979a)*

Mode Quantitatif	Mode Qualitatif
Primauté de la méthode	Primauté du sujet
Manipule des données numériques	Manipule des données langagières
Généralisabilité	Contextualisation
Prédiction	Interprétation
Explication causale	Compréhension

Au-delà de l'accord de principe sur l'utilité respective des méthodes quantitatives et qualitatives dans le champ des IHMs, il convient d'en étudier les propriétés respectives afin de guider leur utilisation convenable en fonction d'un contexte spécifique de recherche. Certaines caractéristiques qui opposent généralement ces deux méthodes sont artificielles, comme l'opposition classique des données numériques et textuelles. En effet, il est possible de réaliser des études quantitatives sur des ensembles de texte grâce à certaines techniques d'analyses statistiques, comme il est possible d'analyser qualitativement des tableaux composés uniquement de chiffres. Néanmoins, il est possible de discerner une différence fondamentale entre ces deux modes en s'appuyant cette fois sur la théorie de l'information de Shannon (1948). Analysons ici ces deux méthodes sous cet angle. Pour la méthode qualitative, l'objectif principal de la recherche est de fournir une description complète et détaillée du sujet de recherche. Elle part d'un point, et tente de saisir toutes les relations autour de l'objet d'étude. Elle cherche à comprendre le sujet dans sa globalité, de tracer son écosystème complet. C'est une recherche dont le nombre d'objets traités est très important. La recherche est riche mais également un peu fragile car chaque observation est analysée distinctement. De ce fait, il y a peu de recoupement, et le mode de recherche, moins structuré, a pour conséquence une exploration non systématique du champ (observation ouverte, entretien non directif). De l'autre côté, la recherche quantitative collecte des données de manière beaucoup plus systématique et agrège, recoupe et triangule massivement les quelques entités d'intérêt choisies par le chercheur. Les données récoltées sont donc beaucoup plus fiables et précises, mais sont, en contrepartie, bien moins nombreuses : en général, quelques-unes par étude. La recherche est donc beaucoup plus pauvre en informations, mais celles-ci sont plus fiables, car la redondance dans la collecte des données permet l'utilisation de procédures statistiques, à la base de la doctrine méthodologique du chercheur quantitatif. En reprenant le dilemme fidélité-bande passante de Cronbach (1960), il est possible de placer les méthodes qualitatives du côté de la « *Bande passante* » et les méthodes quantitatives plutôt du côté de la « *Fidélité* ». Bien entendu, la tendance d'une méthode à tendre

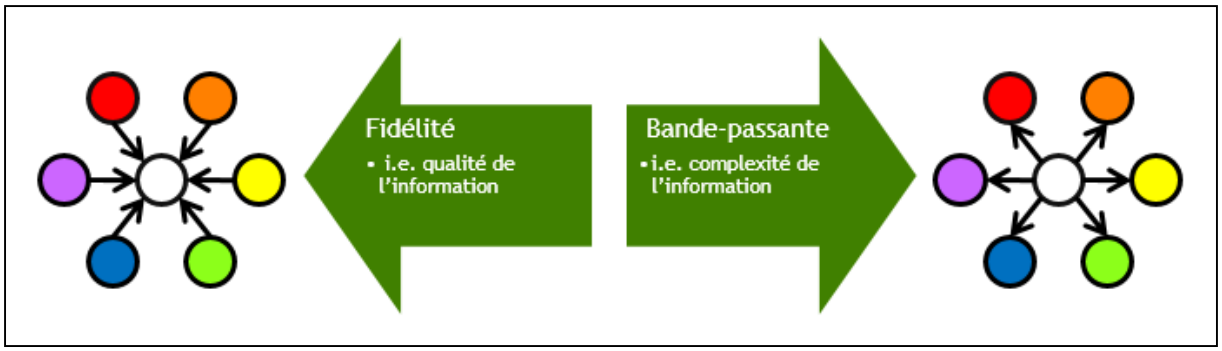


Figure 17 – Continuum Fidélité Bande passante

vers une meilleure fidélité ou à tendre vers une plus grande bande passante (ou richesse d'informations, exhaustivité, …) est à placer sur une échelle continue, plutôt qu'à partir d'une catégorisation fictivement binaire (Figure 17).

Nous verrons que le choix et le développement des méthodes d'évaluation en IHM se sont déplacés sur le continuum fidélité-bande passante au cours du temps. De même, d'autres caractéristiques des méthodes d'évaluation ont été dictées en fonction des dogmes de l'époque, des besoins du marché ou des ressources disponibles. Schématiquement, nous pouvons découper l'histoire des évaluations des IHM en trois paradigmes, dont l'apparition fut successive (Harrison, Tatar, & Sengers, 2007). Néanmoins, il faut avoir à l'esprit que ce découpage est plus pédagogique que réel, car l'évolution a été progressive et les différents paradigmes d'évaluation continuent de coexister. Il est important de détailler les causes inhérentes à la mise en place de ces paradigmes, car ce sont elles qui ont orienté le développement des méthodes d'évaluation au fils du temps, et qui les développeront encore dans le futur. Le chapitre suivant détaillera donc cette évolution pour finir sur le chantier actuel de l'évaluation en IHM.

Paradigme 1 – Ingénierie et facteur Humain

Pour comprendre les paradigmes d'évaluation d'une époque, il faut comprendre quels sont les acteurs en jeu, leurs besoins et le contexte global qui influe sur l'évaluation à un moment donné. Les premières évaluations dans le domaine commencent à la fin des années 40 avec la création du premier vrai ordinateur, l'ENIAC. À cette époque, les premiers ordinateurs étaient en quelque sorte des produits clé en main, sans grande possibilité technique, avec des coûts élevés, une grande complexité fonctionnelle et structurelle, et impliquant souvent un apprentissage de la part des opérateurs (Brangier & Bastien, 2010). De ce fait, le domaine de l'informatique était réservé essentiellement aux spécialistes. Les seuls utilisateurs et évaluateurs étaient donc les ingénieurs et d'autres corps de métier hautement qualifiés. L'intérêt de telles machines était la réalisation de calculs complexes sur de grandes quantités de données. Mais la faiblesse de ces mégastructures était leur fiabilité. De nombreux tubes sous vide tombaient constamment en panne et il fallait les remplacer pour pouvoir continuer à utiliser la machine. Les préoccupations de l'époque ont été la minimisation des temps de dysfonctionnement et la facilité de remise en état du système. L'évaluation d'un tel système se basait donc naturellement sur sa fiabilité de fonctionnement et sa facilité de maintenance (Pollard, 1951). La qualité première d'un

ordinateur était donc sa fiabilité, qui était évaluée principalement par le temps qu'il pouvait fonctionner sans s'arrêter (Kaye & Sengers, 2007).

Puis, au fur et à mesure que les ordinateurs sont devenus plus stables et performants, la programmation est devenue une activité centrale. Dans les années 50-60, de nouvelles techniques d'interactions ont vu le jour (cartes perforées, bandes magnétiques, puis claviers) et ont facilité le développement des premiers langages de programmation. Ces avancées ont permis le développement d'une expertise spécifique qui a amené de nombreux ingénieurs à devenir de véritables informaticiens. À ce moment, la complexité des calculs effectués a dépassé la capacité technique de tels systèmes. De ce fait, il y a eu une évolution du modèle de qualité des systèmes informatisés : un bon système devait à la fois être fiable et puissant techniquement. Il en allait de la viabilité économique de telles machines. Les évaluations à l'époque ont donc consisté à mesurer la capacité d'un système à traiter rapidement un volume important de données sans tomber en panne. Le modèle d'évaluation etait donc, jusqu'à la fin des années 50, complètement technique. Les évaluations étaient réalisées par les informaticiens, pour les informaticiens. Néanmoins, nous voyons apparaître à cette époque les prémices d'un parti pris différent, notamment avec Grace Hooper, militant pour une « *libération des utilisateurs informatiques* » (Hopper, 1952). Véritable pionnière dans le domaine, elle s'attacha à développer divers outils −comme des langages, compilateurs et sous-routines− pour faciliter l'interaction « *programmeur-ordinateur* » (Grudin, 2012a). Pour elle, son rôle a été de « *rendre la liberté aux mathématiciens de faire des mathématiques* » (Hopper, 1952). Malheureusement, ses accomplissements furent sous-évalués et marginalisés par les ingénieurs informaticiens de son époque.

C'est dans les années 60 que le champ de l'interaction Homme machine commença véritablement à se développer par l'incorporation de nouveaux acteurs. Après la Seconde Guerre mondiale, les termes « *d'ingénierie humaine* », « *facteur humain* » ou « *ergonomie* » commencèrent à être popularisés. Ces ergonomes, initialement formés à l'école du Taylorisme industriel ou à l'étude des systèmes de contrôles complexes (tels que dans les secteurs de l'aviation militaire), ont tout naturellement rejoint le domaine des IHM. L'amélioration des boutons, leviers et affichages représentait une extension logique du champ des facteurs humains. Le premier papier IHM, « *Ergonomics for a Computer* » fut écrit par un britannique, Brian Shackel (1959). Il s'agissait de résoudre le problème de l'interface homme-ordinateur, c'est-à-dire d'étudier la compatibilité entre les caractéristiques matérielles et logicielles de l'ordinateur et les caractéristiques physiologiques et mentales de l'opérateur humain. Ainsi, les premières études ont abordé les questions du dimensionnement des ordinateurs et de leur adaptation aux caractéristiques physiques et physiologiques humaines (Brangier & Bastien, 2010). Cette notion de couplage homme-machine était au centre des recherches et l'objectif à l'époque était d'en repérer les dysfonctionnements, afin de les résoudre de manière pragmatique et souvent a-théorique. L'intervention de l'ergonome se déroulait généralement à la fin du processus, son action porta le titre d'ergonomie de correction. De plus, les ergonomes informatiques étaient assez marginalisés, car on ne considérait pas que le travail sur ordinateur fût épuisant, contraignant ou dangereux. Leur enjeux étaient donc perçu comme léger en comparaison de ceux travaillant dans l'industrie lourde ou étudiant les travailleurs en conditions extrêmes, telles que dans la sidérurgie ou les mines.

Les progrès et orientations générales dans les IHM à cette époque se basaient sur des visions grandioses du monde plutôt que sur des études de marchés (marché qui n'existait tout simplement pas). Ces visions étaient portées par de grandes figures du domaine, qui allaient chacune s'efforcer de réaliser:

- **La symbiose Homme-Machine *(*J.C.R Licklider*)*** : En 1960, le psychologue américain Licklider présenta sa vision de la symbiose Homme-Machine (*Man-Computer Symbiosis*) qui va préfigurer l'informatique interactive: « *Notre espoir est que, dans un certain nombre d'années, le cerveau et l'ordinateur soient couplés très étroitement. Grâce à ce partenariat, on pourra penser comme jamais auparavant, et les données seront traitées d'une façon incomparable à aujourd'hui* ». Cependant, pour mettre en place un tel partenariat, il émit la condition qu'il faudra d'abord multiplier les occasions de rencontre entre l'homme et les machines (développement du temps partagé) et rendre plus accessible les opérations qui se produisent en son sein (développer les interfaces d'entrée et de sortie). Il financera également les premières recherches sur les réseaux informatiques et notamment ARPAnet, l'ancêtre d'Internet. De plus, l'article qu'il écrivit avec Robert Taylor « *The Computer as a Communication Device* » (Licklider & Taylor, 1968) prédira l'utilisation de réseaux d'ordinateurs pour des groupes ayant les mêmes centres d'intérêts, ainsi que la collaboration sans critères de localisation.
- **Augmenter l'intelligence humaine (Douglas Engelbart)** : En 1962, l'ingénieur américain Engelbart présente sa vision dans le papier « *Augmenting Human Intellect: A Conceptual Framework* » (Engelbart, 1962). Il écrivit : « *par augmenter l'intelligence de l'homme, nous entendons accroître la capacité d'un homme à aborder une situation problématique complexe, gagner en compréhension selon les besoins, et en tirer des solutions pour résoudre les problèmes. L'un des objectifs est de développer de nouvelles techniques, procédures et systèmes qui permettront de mieux adapter les aptitudes informationnelles des personnes aux besoins, problèmes et de permettre le progrès de la société* ». Engelbart est l'un des premiers à avoir compris très tôt le potentiel d'innovation lié à l'informatique interactive. Il a construit un programme de recherche sur sa vision et l'a mis en œuvre les années qui suivirent avec brio. Son programme visait à faire progresser notre intelligence collective pour répondre à l'accélération de la complexité du monde et de l'urgence des problèmes qui se posent à nous. Sa vision était large, son approche systématique. L'ordinateur était pour lui un moyen d'augmenter nos capacités intellectuelles, comme l'avait fait l'écriture ou l'imprimerie. Il souhaitait l'adapter à nos capacités, mais également adapter nos pratiques à ce qu'il permettrait de nouveau. En effet, il faisait la distinction entre l'adaptation de l'ordinateur aux pratiques existantes dans le but d'automatiser les tâches répétitives (demi-augmentation), et celles qui créent radicalement de nouvelles pratiques (et qui nous transforme radicalement). Ses recherches se sont ainsi intéressées à la visioconférence, la téléconférence, au courrier électronique, aux «

Figure 18 – Premier prototype de la souris ©SRI

f*enêtres* » et aux liens hypertexte, bien qu'il soit surtout connu de nos jours pour avoir inventé la souris d'ordinateur (Figure 18).

- **Un monde interconnecté (Ted Nelson) :** durant les années 60, le sociologue Ted Nelson inventa le terme *Hypertext* et définit ce que sera Internet, aux côtés de Vint Cerf (le père de l'Internet), Tim Berners-Lee (créateur du World Wide Web) et Douglas Engelbart. En 1965, dans son papier « *A File Structure for the Complex, the Changing and the Indeterminate* », il écrit « *Laissez-moi vous introduire le mot Hypertexte pour définir un corpuscule d'écrits ou d'images interconnectés d'une manière si complexe qu'elle ne peut pas être représentée convenablement sur le papier* » (Nelson, 1965). Il continuera de militer pour la démocratisation de l'informatique construite sur un monde d'informations interconnectés et accessibles à tous (Nelson, 1973).

Ainsi, l'informatique a été construite dans les années 60, non par des études utilisateurs ou de marché, mais sur la base de quelques visionnaires qui ont contribué pendant plus de 40 ans au bouleversement de notre monde par les technologies de l'information. Ce n'est pas une coïncidence si Alan Kay, une des personnalités fortes du XEROX PARC dans les années 70, déclara que : « *le meilleur moyen de prédire le futur est de l'inventer* ». L'évaluation des facteurs humains aux contacts des systèmes informatiques y était encore informelle et sous la tutelle des figures dominantes : celles de l'ingénieur et de l'informaticien. Cela était à l'époque tout naturel, car ces derniers étaient en même temps les créateurs et les utilisateurs de ces systèmes. Néanmoins, la sortie de l'informatique des laboratoires changea radicalement la donne.

Paradigme 2 – La révolution cognitive

L'ordinateur d'entreprise et la recherche utilisateur

C'est au milieu des années 60 que les premiers ordinateurs incorporent les premières entreprises. Dans ce contexte, l'informatique y avait un objectif de centralisation des données et de traitement de masse dans un marché en pleine expansion. L'amélioration de ces systèmes dans les années 70, en permettant l'interrogation de données sur des terminaux séparés, conduisit à l'informatisation des postes administratifs. Très rapidement, le prix du matériel baissa et les périphériques d'interaction devinrent des éléments de plus en plus vitaux. James Martin, dans le premier livre sur les IHM largement lu, au sujet de cette transition, écrivit : « l*e terminal ou la console de l'opérateur, à la place d'une considération périphérique, deviendra la queue qui remuera le chien entier… L'industrie informatique sera forcée de se concentrer davantage sur les usages des gens que sur les boyaux des ordinateurs* » (Martin, 1973). L'introduction progressive des systèmes à temps partagé permit à de nombreux opérateurs d'utiliser les capacités de l'ordinateur central en même temps. De ce fait, l'introduction des terminaux passifs pour la saisie à la source, puis pour l'interrogation se multiplièrent.

Le développement des interfaces graphiques et l'utilisation de plus en plus grande des ordinateurs pour des tâches autres que la programmation aura pour conséquence d'augmenter considérablement le nombre d'utilisateurs « *non-spécialistes* » (MacDonald & Atwood, 2013). Cela força progressivement les évaluateurs à se préoccuper davantage de la performance des

utilisateurs plutôt que de la vitesse des systèmes informatiques. Néanmoins, l'objectif de rentabilité des entreprises, a contrario des laboratoires, ne s'est pas fait sans difficulté. Même si la principale préoccupation de nombreuses entreprises était simplement de paraître moderne, le désir de rendre rentable l'investissement de plus d'un demi-million de dollars pouvait enchaîner un manager à un ordinateur, le condamnant à gérer sans interruption le flux permanent des données (Grudin, 2012a). De plus, l'opérateur administratif ou le manager ne voyait pas dans l'utilisation élitiste d'un système une démonstration d'une expertise à faire valoir. Il y voyait plutôt l'annonce certaine d'un mal de tête en perspective. C'est pour cela que des approches sociotechniques participatives (Björn-Andersen & Hedberg, 1977; Mumford, 1971) ont commencé à se développer en réponse aux difficultés et résistances. Néanmoins, les acteurs les plus nombreux à s'être vraiment engouffrés dans le champ de l'évaluation furent les psychologues expérimentaux.

En 1967, Ulric Neisser publie *« Cognitive Psychology* », l'ouvrage qui marqua l'explosion de la psychologie cognitive. Depuis lors, la révolution cognitive envahit les IHM, d'autant plus que la théorie derrière ce nouveau paradigme restera très liée à celle de l'informatique. Son idée centrale est que le système de traitement de l'information humain est fortement analogue à celui du traitement du signal informatique. De ce fait, la tâche première du domaine des IHM est de permettre la communication entre l'homme et la machine (Card, Newell, & Moran, 1983). Ainsi, de nombreux psychologues cogniticiens, dans les années 70, puis massivement dans les années 80, généralisèrent les études utilisateurs en laboratoire afin d'améliorer la productivité des opérateurs informaticiens (Gaines, 1985). Ils se basèrent sur des indicateurs de performance tels que l'efficacité, l'efficience, la précision, le nombre d'erreurs ou encore la facilité d'apprentissage. La satisfaction est également étudiée, mais plutôt sous le prisme de l'acceptation, car le choix des outils informatiques était encore imposé dans l'entreprise. Ces études fondèrent les bases méthodologiques et théoriques du domaine, que Shneiderman (1980) appellera la « *psychologie logicielle* », et dont le but sera de *« faciliter l'utilisation humaine des ordinateurs* ». Il s'agissait également de prouver l'utilité des approches comportementales pour améliorer la compréhension de la conception logicielle, la programmation, et l'utilisation de systèmes interactifs, afin de motiver et guider les concepteurs à considérer les caractéristiques humaines (Carroll, 1997). Le caractère scientifique de l'étude représentait à l'époque un argument d'autorité fort utile. En effet, des études expérimentales poussées étaient souvent nécessaires pour convaincre par A+B les ingénieurs de changer un système, qui coûtait encore à l'époque plusieurs centaines de milliers de dollars en moyenne. De plus, à cette période −et pendant un certain temps− ce que pouvaient penser les futurs utilisateurs de leurs systèmes les concernaient que peu. Même bien plus tard, Hammer (1984) statuait encore que l'utilisabilité était de second ordre et que ce qui importait pour une application de bureau était : (1) les fonctionnalités, (2) les fonctionnalités, (3) rien d'autre, (4) les fonctionnalités et (5) ce qui reste… La culture de l'ingénieur a toujours considéré le fait de surmonter les difficultés techniques comme une fierté car cela démontre une certaine expertise acquise. La plupart des utilisateurs tout venants cherchent au contraire à les éviter.

Dans ce contexte, les psychologues expérimentaux assumèrent deux rôles : (a) produire des descriptions générales de l'interaction humaine avec des systèmes informatisés, qui peuvent être synthétisées sous forme de guideline pour les concepteurs ; (b) vérifier directement

l'utilisabilité de systèmes informatiques après leur développement (Carroll, 1997). Ces pratiques suivaient une vision top-down de la constitution de la connaissance avec une volonté forte de généralisation des savoirs. Elle sous-entend un cycle de conceptions en cascade, c'est-à-dire linéaire, avec une phase de spécification, de conception, de validation et de maintenance (Royce, 1970). En amont, elle orientait la conception dans la phase de spécification à partir de connaissances théoriques ou de modèles comportementaux prédictifs (ex : GOMS ; Card, Moran, & Newell, 1983). Puis, elle contrôlait le système durant la phase de validation par une évaluation empirique à caractère soit expérimental, soit analytique.

De nombreux progrès seront réalisés par l'incorporation et l'application de savoirs fondamentaux issus de la psychophysique[18] (temps de latence acceptable, taux de rafraichissement, luminosité, principe de regroupement visuel, …), de la psychologie cognitive (limite attentionnelle, mémorisation, apprentissage, …), et du développement de modèles de traitement de l'information complets, comme le modèle du processeur humain (Figure 19 ; Card et al., 1983) qui permit avec GOMS de simuler la performance d'un utilisateur dans un environnement informatique donné. D'autres études, plus appliquées, portèrent sur l'efficacité de la souris (English, Engelbart, & Berman, 1967), la complexité relative des constructions syntaxiques dans les langages de programmation (Sime, Green, & Guest, 1973) ou encore de

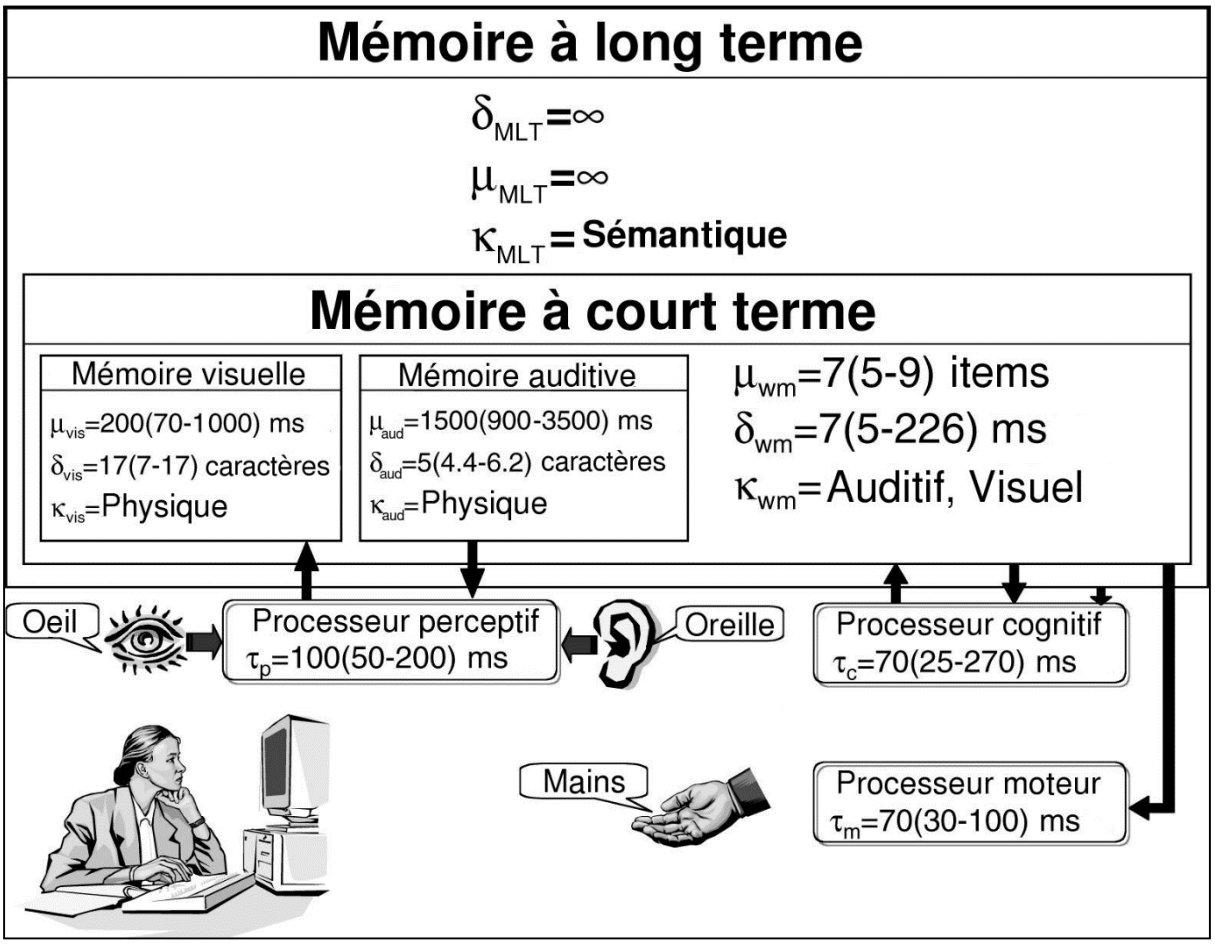


Figure 19 – Le modèle du processeur humain *(Jacko & Sears, 2002)*

[18] La psychophysique est une branche de la psychologie expérimentale qui cherche à déterminer les relations qui existent entre un stimulus (visuel, auditif ou tactile) et la perception qu'on en a.

l'utilité mnémotechnique des noms de variables et des commentaires pour les lignes de code lors de la programmation (Weissman, 1974).

L'ordinateur personnel et la quête de l'utilisabilité

Jusqu'à la fin des années 70, les seules personnes pouvant interagir avec les ordinateurs étaient encore des professionnels des systèmes informatiques ou des amateurs passionnés. Cela changea brusquement lors de l'introduction de l'ordinateur personnel au début des années 80. Le PC (*Personal Computer*) incarna une solution clef en main, peu coûteuse, peu encombrante et beaucoup plus facile à utiliser qu'auparavant, notamment par la démocratisation du mode d'interaction WIMP (Windows, Icon, Mouse, Pointer). À partir de ce moment, tout individu devint ainsi un utilisateur potentiel et la nécessité de fournir des logiciels faciles à prendre en main devint un besoin économique vital (Carroll, 2013). En effet, pour rester compétitif dans le milieu du développement logiciel, la norme consista à proposer des produits utilisables sans formation nécessaire, au point où de nombreuses entreprises se mirent à embaucher un grand nombre de psychologues expérimentaux pour contrôler l'utilisabilité des produits avant leur mise en vente sur le marché. De même, le besoin de développer une « *psychologie de l'usage volontaire* » émergea dans les entreprises où l'utilisateur pouvait résister, saboter ou quitter le travail dans le cas d'outils trop contraignants à utiliser (Grudin, 2012a). De plus, les gens commencèrent à distinguer finement entre les développements exploratoires poussés par la technologie (« *Techno Push* »), et les développements prenant l'utilisateur comme l'arbitre final et dans lequel la convivialité est vérifiée empiriquement (Carroll, 1997). La première conférence CHI, en 1983, confirma cette tendance. Les psychologues cognitivistes dominèrent le programme, même si la « *Human Factor Society* » co-sponsorisa encore la conférence. L'étude de l'usage novice y est particulièrement prépondérante. En effet, l'on se rendra compte très rapidement que la qualité de la première expérience d'utilisation d'un système informatique par un utilisateur volontaire (en opposition avec l'utilisateur forcé professionnellement) est vitale pour son adoption. Leurs études intéressèrent particulièrement les entreprises de développement logiciel visant le marché gigantesque du grand public. De plus, on remarqua que le choix de l'utilisateur novice était particulièrement judicieux, car dans un environnement où de nouveaux produits sont constamment créés, l'utilisateur peut être vu comme un novice perpétuel qui doit constamment s'adapter. Enfin, dans la conjoncture de l'époque, où de plus en plus de personnes se mettaient à l'informatique, le nombre de novices augmentait invariablement d'une année à l'autre.

Ce nouveau contexte, constitué par (i) l'avènement du marché de l'ordinateur personnel, (ii) la focalisation sur l'utilisateur novice, et (ii) la compétition de plus en plus rude entre les éditeurs logiciels pour capter leurs faveurs, questionna un grand nombre de pratiques de conceptions et d'évaluations des systèmes informatiques qui avaient prévalu jusqu'alors. Les axiomes de base de l'organisation du travail d'évaluation se trouvèrent être alors remis en cause (Carroll, 1997) :

(a) Le modèle de développement en cascade apparut comme inefficace ou alors seulement dans des projets longs et à grande échelle. Or, le développement logiciel, depuis les années 80, se réalisait souvent dans de petites équipes distribuées et avec des cycles de développement compressé en moins d'une année. Ces différences s'expliquaient par un développement plus souple, flexible et rapide que celui du développement matériel.

(b) La division du travail des psychologues logiciels dans les laboratoires s'est trouvée être contre-productive (Carroll, 1997). En effet, les chercheurs, souvent en laboratoire universitaire, avaient pour rôle de créer des guidelines pour les psychologues en milieu industriel. Plusieurs problèmes se posaient alors. D'une part, il existait un décalage important entre la situation réelle et celle testée en laboratoire : les tâches étaient simplifiées, les alternatives de conception exagérées, et les sujets peu représentatifs par rapport aux véritables utilisateurs. De ce fait, les psychologues dans l'industrie avaient du mal à utiliser ces résultats dans un contexte souvent beaucoup plus riche et, surtout, très diffèrent. De plus, ils avaient du mal à vérifier l'utilisabilité des systèmes en fin de cycle, car les méthodes d'évaluation qu'ils utilisaient (expérimentation formelle dans l'objectif de mesurer les différences entre plusieurs alternatives) étaient très peu informatives, longues à réaliser et coûteuses. De ce fait, les spécialistes en facteur humain dans l'entreprise étaient assez couramment vus comme des obstacles bureaucratiques bloquant l'héroïsme des développeurs…

Cela a mené à la naissance de nombreux points de discorde épistémologiques, stratégiques et méthodologiques dans le champ d'étude des IHM. Les recherches sur la modélisation des performances utilisateurs sont de bons exemples de travaux qui ont fait débat à l'époque sur le rôle de la recherche en IHM (Card, Moran, & Newell, 1980a, 1980b; Card et al., 1983; Newell & Card, 1985). Ces derniers étaient partisans d'une construction méthodologique et théorique plus rigoureuse dans le domaine. Ils voyaient donc l'avènement de la modélisation utilisateur comme une grande avancée de la psychologie cognitive. Pour eux, ces fondements scientifiques en étaient renforcés : Newell et Card (1985): « *les facteurs humains classiques ... ont toutes les caractéristiques qui font d'elle une discipline de seconde classe. (Notre approche) évite la continuation du rôle classique des facteurs humains (en transformant) la psychologie des interfaces en une science dure* » (Newell & Card, 1985, p. 221). Ces modèles permirent d'intégrer explicitement de nombreuses caractéristiques utilisateurs pour produire des prédictions sur des tâches réelles. Ils furent appliqués avec succès, notamment lors de l'étude des opérateurs téléphoniques (Gray, John, & Atwood, 1992; Gray, John, Stuart, Lawrence, & Atwood, 1990). Néanmoins, la modélisation des performances humaines attira un nombre modeste d'adeptes en IHM. De nombreux autres acteurs, au contraire, militèrent pour une recherche plus souple et pragmatique, en accord avec les attentes des entreprises. En effet, la modélisation démontra son utilité pour les domaines où la performance était primordiale. Or, c'était plutôt la facilité d'apprentissage qui était au cœur des préoccupations dans les années 1980. De plus, cette technique nécessitait l'assistance d'experts et, de préférence, sur un système bien défini, car les prévisions temporelles dépendent des caractéristiques physiques et perceptives de l'interface (Olson & Olson, 1990).

Ainsi, une séparation progressive se dessina entre les facteurs humains et les IHM. D'un côté, les acteurs de l'ergonomie et des facteurs humains continuèrent à se baser sur les méthodologies traditionnelles. Ils concentrèrent leurs efforts sur les opérateurs (utilisateurs non volontaires) et privilégièrent la méthode expérimentale pour améliorer leurs performances. Ils continuèrent de répondre aux besoins des organismes gouvernementaux, de l'armée, de l'aviation et des industries lourdes. L'accent fut mis sur les utilisateurs qualifiés et formés. En effet, pour la réalisation de tâches routinières, même de petits gains d'efficacité dans les opérations individuelles peuvent engendrer de grands avantages au fil du temps, ce qui justifie l'effort

d'apporter des améliorations qui pourraient ne pas être remarquées par les utilisateurs non professionnels (Grudin, 2012a). De plus, des tests de validation poussés sont indispensables dans les industries sensibles, où nulle erreur n'est permise, tels que dans l'aérospatial ou le nucléaire. Les cycles de conception de système plus complexe ou nécessitant un développement matériel propre sont également plus adaptés à l'application des méthodologies traditionnelles, car ils sont généralement beaucoup plus longs.

De l'autre côté, les acteurs IHM se concentrèrent sur le marché de l'usage volontaire, que ce soit dans le monde de l'entreprise ou en dehors. L'évaluation de l'utilisabilité devint la norme, surtout auprès des utilisateurs novices, première cible du nouveau marché du logiciel. Pour rester compétitifs, ils mirent au point tout un nombre de méthodes d'évaluation et de conception, plus souples et plus holistiques que les précédentes. On demanda d'elles d'être efficientes, et non plus seulement valides. En effet, il fut constaté comme plus utile d'avoir une information non entièrement certaine, obtenue à temps, qu'une information plus solide, trop tard. Cela était d'autant plus nécessaire que l'instabilité causée par la rapidité du changement technologique sapa l'élaboration de savoir théorique de plus en plus vite. En effet, les chercheurs avaient à peine le temps de poser les fondations théoriques d'un travail de recherche qu'une vague d'innovations rendait le tout obsolète. On peut penser aux chercheurs des années 80 dont le travail reposait sur les moyens d'optimiser le langage par ligne de commande. Le succès commercial de l'interface utilisateur graphique (GUI), en 1985, rendit ce projet sans intérêt, au grand dam de ceux qui espéraient que leur travail servît de fondation pour la recherche et le développement futur (Grudin, 2012). Ces nouveaux acteurs, « *croyant en l'expertise et non aux expérimentations* » (Nielsen, 1984 ; cité par Kaye & Sengers, 2007), vont développer un ensemble de méthodes qu'ils vont rassembler sous la bannière de *« l'ingénierie d'utilisabilité* » (« *Usability Engineering* »).

L'accélération du marché logiciel et la création de l'ingénierie d'utilisabilité

En moins de dix ans, l'ingénierie d'utilisabilité devint le nouvel étendard sous lequel diverses méthodologies furent formalisées et approfondies durant les années 80 et 90. De plus, l'organisation du travail de l'ergonome changea et s'adapta aux nouvelles contraintes du milieu. On abandonna la conception en cascade, peu efficace dans le contexte très réactif de la création logicielle, dans les années 80. Ainsi, la diffusion de guidelines abstraites en début de projet et suivie de validations strictes en fin de projet laissa la place à un processus de développement complet, intératif et centré-utilisateur.

Un des grands bouleversements du rôle des ergonomes fut l'extension de son intervention. Les évaluations furent réalisées plus tôt dans le projet et plus régulièrement. Les utilisateurs réels furent sollicités tout au long du cycle et leurs feedbacks permirent de rediriger la conception bien plus tôt, lorsqu'il reste encore des marges de manœuvre et avant que les changements nécessaires deviennent trop coûteux. De même, l'ergonome ne fut plus seulement là pour dispenser des recommandations, en restant à l'écart de la conception. Chapanis (1996) écrit ainsi : «*[Les manuels sur les facteurs humains] fournissent de nombreuses règles et recommandations générales sur les exigences des utilisateurs, fondées sur des résultats de recherche. Règles et exigences sont écrites avec l'hypothèse implicite, parfois explicite, que les*

concepteurs les liront et qu'ils en déduiront comment concevoir des objets adaptés aux capacités et limites humaines. Le problème de cette approche est que, en gros, elle ne marche pas. Les ingénieurs, les concepteurs et les programmeurs ne lisent pas nos manuels, ne comprennent pas nos règles et recommandations dans le cas où ils les liraient, et ne savent pas comment concevoir de façon à satisfaire nos règles dans le cas où ils les liraient et tenteraient de les suivre. Il n'y a pas de raison pour qu'ils y parviennent. Nous ne devrions pas attendre des concepteurs qu'ils accomplissent un travail pour lequel nous avons été formés et pas eux » (Chapanis, 1996, traduction de Falzon, 2005). Scriven (1983), ne pensait pas autrement. Pour lui, « *le plus grand échec d'un évaluateur est de procurer simplement des informations aux décideurs, de "passer le sceau" (pour le jugement final) aux non professionnels* » (p. 248). L'ergonomie en conception (dont les méthodes porteront le nom d'ergonomie de conception) donna un rôle actif à l'ergonome, celui d'accompagner les développeurs tout au long de la création du produit.

C'est également pour les ergonomes l'opportunité d'endosser une responsabilité plus large. Cela signifie qu'ils ne durent plus seulement revêtir le rôle du gardien, garant d'une adéquation humain-machine satisfaisante, mais également celui de porte-parole des utilisateurs, dont les besoins commencèrent à s'exprimer et à se confronter aux aprioris des concepteurs. Le livre de Cooper (1999) « *The Inmates are Running the Asylum* » (« *Les détenus dirigent l'asile* » (Figure 20), en résuma les enjeux : « *Imaginez, à un rythme terriblement agressif, que tout ce vous utilisez régulièrement est remplacé par une technologie informatique. Pensez à votre téléphone, aux appareils photos, aux voitures automatisées et programmées par des gens qui, dans leur empressement à accepter les nombreux avantages de la puce de silicium, ont abdiqué leur responsabilité de rendre ces produits faciles à utiliser. Ce n'est pas une exagération, c'est une réalité. Nos vies sont de plus en plus centrées sur les caprices, traits d'esprits, décisions et catastrophes de l'industrie high-tech. Et ces développeurs logiciels, matériels et technologiques ne pensent pas comme nous. Malgré les apparences, les dirigeants d'entreprises ne sont pas ceux qui contrôlent le monde high-tech - ce sont les ingénieurs qui animent le show. Nous avons laissé les détenus diriger l'asile (...) Nous avons besoin que ces créateurs d'innovations toujours "plus rapides, plus fortes, meilleures" améliorent véritablement notre travail et notre vie - pas seulement nous rendre fou. Nous avons besoin que la technologie fonctionne de la même manière que les gens pensent. Nous avons besoin d'une révolution pour rétablir notre santé mentale* » (Cooper, 1999, page de garde). En effet, quand l'ergonome est absent du processus de conception, ce sont les concepteurs de systèmes qui doivent prêter aux utilisateurs des caractéristiques afin d'anticiper

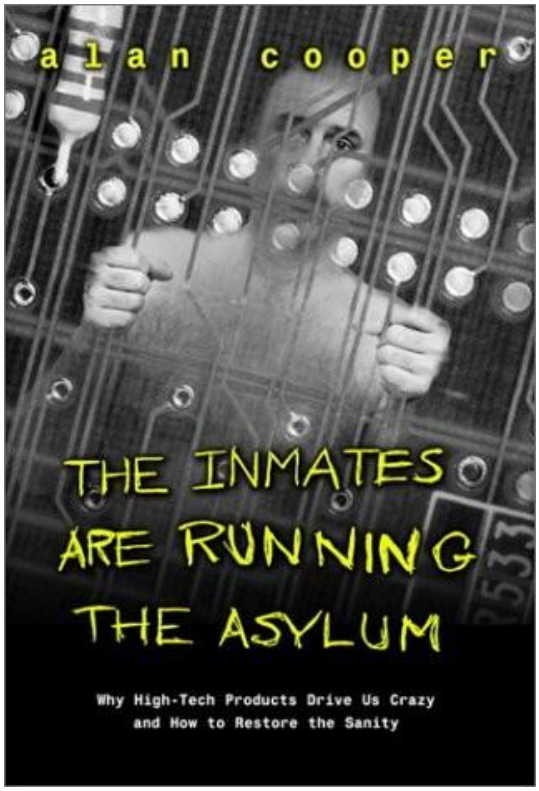


Figure 20 – Page de garde du livre « Les détenus dirigent l'asile » (Cooper, 1999)

les usages des dispositifs qu'ils conçoivent. Ils se basent pour cela sur deux stratégies. Dans la première, les concepteurs se basent sur une réflexion basée sur le « *bon sens* » ; dans la seconde, les concepteurs s'identifient aux usagers. Ces deux stratégies sont évidemment limitées. La première parce qu'elle prête *a priori* certaines caractéristiques aux utilisateurs ; la seconde parce que les concepteurs sont des experts dans leur domaine, ce qui n'est pas le cas des futurs usagers. Or, de telles anticipations, qui seront matérialisées dans l'artefact ou le système, seront autant de sources de difficultés, d'incompréhensions ou de décalages pour les personnes, car reposant sur des modèles faux ou insuffisants. La conception doit donc inclure l'utilisateur réel dans son développement, car le recueil de leurs feedbacks est un garde-fou qui permet à un projet de ne pas s'égarer, voire au contraire, de trouver son chemin : c'est la conception centrée sur l'utilisateur (ou conception orientée utilisateur). C'est une démarche de conception où les besoins, les attentes et les caractéristiques propres des utilisateurs finaux sont pris en compte à chaque étape du processus de développement d'un produit. Elle donnera lieu à une norme en 1999, l'ISO 13407. Elle se développa jusqu'à nos jours et son utilisation démontrera, entre autres, une délimitation plus précise des objectifs du projet (Jacob Nielsen, 1993); une meilleure acceptabilité du système (Damodaran, 1996) ; des avantages concurrentiels conséquents grâce à des produits plus ciblés (Cooper & Kleinschmidt, 2000) ; ou encore l'abandon de fonctions coûteuses jugées non utiles ou non souhaitées par les utilisateurs (Anastassova, 2006). De plus, le feedback utilisateur permit de faire émerger des usages auxquels nous ne pensions pas à priori et d'ainsi d'alimenter en retour le développement de nouvelles solutions technologiques.

La troisième caractéristique forte de ce renouveau méthodologique fut basée sur un compromis nécessaire. En effet, évaluer plus tôt, plus souvent, et plus de choses (au-delà du cognitif) ne se fit pas sans poser de gros problèmes logistiques : celui des ressources nécessaires. Nous avons vu avec le dilemme fidélité bande-passante, qu'à un niveau de ressources égal, nous devrons choisir entre l'exhaustivité (bande-passante) ou la fiabilité de l'évaluation. Ici, il est clair que les évaluations, plus riches et régulières, limitent les ressources disponibles pour chacune. Elles sont donc mécaniquement moins solides. Les méthodes d'évaluation expérimentales, au contraire, font le choix délibéré de porter le regard sur peu d'objets d'étude en même temps, tout en y injectant beaucoup de ressources. Ces méthodes sont donc peu informatives (elles informent sur peu d'entités à la fois), longues à réaliser mais suffisamment fiables pour participer à la construction durable d'un champ académique. Elles contribuent à la progression d'une communauté, dont les membres peuvent rattraper leurs retards respectifs grâce au partage des connaissances. Par contre, leurs lenteurs méthodologiques permettent difficilement d'innover rapidement. C'est pour cela qu'elles sont rarement utilisées dans le monde de l'entreprise. En effet, au vu de l'accélération de la course technologique et de la concurrence exacerbée, ces dernières ont préféré le recueil riche d'informations incertaines à temps (ex : évaluation formative), plutôt que de quelques d'informations sûres mais tardives (ex : évaluation expérimentale). Le rapport qualité/coût est donc une question au cœur de l'ingénierie d'utilisabilité (Carroll, 1997). Il va de soi qu'une entreprise, dans un contexte de compétition et de forte réactivité, est prête à prendre plus de risques qu'une communauté de recherche.

Néanmoins, il est possible de faire un compromis entre les deux approches au fil du développement. Il s'agit pour cela d'utiliser des méthodes « riches », au début de l'investigation, pour capter le maximum de problèmes ; puis, après réduction du champ des possibles, d'utiliser des méthodes plus « rigoureuses », vers la fin du projet, pour valider les choix réalisés (Figure 21). En effet, la marge de modification du produit est très grande au début du projet, puis diminue lorsque les fonctions techniques du produit sont privilégiées, délimités puis développées. L'ergonome peut donc faire appel à des méthodes plus précises quand le nombre d'éléments à tester diminue. C'est pour cela que l'on utilise couramment des méthodes qualitatives en début de projet (entretien, observation, …) et des méthodes quantitatives en fin de projet (tests utilisateurs, questionnaires, …).

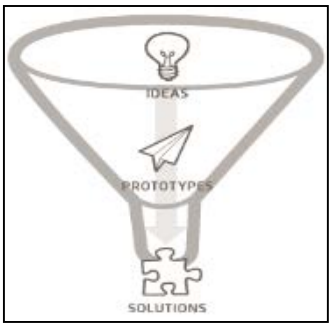


Figure 21 – Entonnoir de Martin (2010) adapté par Bisset *(2012)*

Les nouvelles méthodes d'ingénierie de l'utilisabilité

De nombreuses méthodes nouvelles furent mises au point ou adaptées à partir d'autres disciples afin de couvrir tout le cycle de développement du produit. Il n'est pas question ici de les aborder dans leur intégralité. Néanmoins, il est utile de citer celles qui ont participé à l'évolution de la portée de l'évaluation en IHM afin de comprendre le mouvement qui a animé le domaine.

Méthodes en amont de la conception

Au début du cycle de conception, deux méthodes clefs illustrèrent la volonté d'ancrer l'utilisateur réel très tôt dans la conception : les PERSONAS (Cooper, 1999; Jenkinson, 1994) et l'ANALYSE DU CONTEXTE D'UTILISATION (Wixon, Holtzblatt, & Knox, 1990). Les PERSONAS furent décrits en premier dans le monde du marketing, et plus précisément dans le papier de Jenkinson (1994) « *Beyong segmentation* », afin d'aller plus loin que la segmentation classique du marché. Son approche consista à décrire plusieurs archétypes de clients potentiels sous la forme de personnages imaginaires, mais dont les attitudes et comportements avec une marque spécifique reflètent la réalité. En parallèle, Allan Cooper développa un concept similaire, en 1995, qu'il popularisa dans son livre de 1999. Les personas y sont décrits comme des personnages fictifs qui sont créés pour représenter les différents types d'utilisateurs d'un groupe démographique ciblé, et qui pourrait utiliser un site ou un produit (Cooper, 1999). Utilisée dans le cadre d'un processus de conception centrée utilisateur, cette technique nous permet de prendre en compte les objectifs, les désirs et les limites de l'utilisateur lors de la conception du produit. Ils sont construits sur la base de recherches ethnographiques, d'enquêtes ou d'entretiens. Les noms et les photographies peuvent être fictifs, mais les détails doivent être factuels. L'utilisation de la méthode des PERSONAS offre de nombreux avantages dans le développement de produits. Les PERSONAS sont cognitivement convaincants, car ils mettent une face humaine et personnelle sur des données utilisateur qui seraient restées sinon abstraites et indigestes (Pruitt & Adlin, 2006). Les PERSONAS aident aussi à prévenir le concepteur des deux écueils communs de conception décrits plus haut. (i) Le premier est ce que Cooper appelle « *l'utilisateur élastique* », qui désigne la tendance de différents intervenants à prendre des décisions sur une image de l'utilisateur formée selon leurs convenances. Ici, les PERSONAS aident l'équipe à développer une

compréhension partagée des utilisateurs réels en termes de motivations, capacités et du contexte. (ii) Les PERSONAS permettent également d'éviter la conception « *autoréférentielle* », c'est-à-dire quand la tendance du concepteur ou développeur de projeter inconsciemment leurs propres modèles mentaux sur la conception du produit (et qui est souvent très différent de ceux de la population cible). De plus, les PERSONAS aident les concepteurs à maintenir le cap de la conception sur les cas qui sont les plus susceptibles d'être rencontrés par les utilisateurs cibles et non sur des cas limites qui ne se produisent généralement pas (Cooper, 1999).

L'ETUDE DE TERRAIN est également devenue une approche importante dans les IHM (Whiteside & Wixon 1987). Elle est plus connue sous le nom de CONCEPTION CONTEXTUELLE (« CONTEXTUAL DESIGN » ; Wixon et al., 1990). Elle permet de contrebalancer les biais des tests en laboratoire, qui simplifie la situation réelle dans les tâches ou les situations qu'elles simulent. En effet, l'étude sur le terrain permet d'apporter la lumière sur des faits en arrière-plan du contexte d'utilisation, c'est-à-dire des circonstances d'usages dont les utilisateurs eux-mêmes n'ont pas conscience (Carroll, 1997).

L'analyse en amont peut se compléter par des méthodes plus classiques comme les entretiens, les enquêtes par questionnaire ou par d'autres méthodes encore plus novatrices à l'époque :

- **Le FOCUS GROUP** : importé du marketing, cette méthode consiste en un entretien structuré de groupe qui permet de révéler rapidement et de manière peu coûteuse les désirs, expériences, priorités ou valeurs d'une communauté ciblée (Nielsen, 1997; Bruseberg & McDonagh-Philp, 2002; Klein, Tellefsen, & Herskovitz, 2007).
- **La MODELISATION DE TACHES CENTREES UTILISATEUR** : Ce sont des méthodes qui permettent d'accéder aux représentations propres des utilisateurs sur les tâches à accomplir (Scapin & Bastien, 2001; Scapin, 1988) et qui sont souvent très différentes de celles des concepteurs (Mayhew, 2009). Ces méthodes ont accompagné la tendance centrée-utilisateur des années 80 qui tendait à remplacer la documentation centrée sur le système (Hackos & Redish, 1988);
- **Le TRI DE CARTE** : c'est une méthode d'organisation des contenus qui permet de mettre au point l'architecture d'information d'un système en se basant sur les représentations mentales des utilisateurs. Elle permet de découvrir comment les utilisateurs regroupent des concepts, en leur demandant de trier et de grouper des cartes sur lesquelles sont inscrites des informations décrivant des contenus (Nielsen, 1995; Barrère, Sloim, Bastien, & Mazzone, 2012)

Ainsi, toute une palette d'approches et de techniques a été développée pour favoriser la participation des utilisateurs très tôt dans le cycle de conception. En allant même plus loin, des activités coopératives « peu techniques » ont été suggérées. Elles ont pour objectif de faciliter la collaboration entre utilisateurs (qui apportent l'expertise sur la situation de travail) et les développeurs (qui apportent l'expertise technologique). C'est le courant de la CONCEPTION PARTICIPATIVE (Muller & Kuhn, 1993; Schuler & Namioka, 1993), qui élargit encore les rôles des utilisateurs dans la conception. En effet, dans cette approche, les utilisateurs sont impliqués dans la définition même des objectifs de conception et sont amenés à participer à la planification des premiers prototypes. Néanmoins, deux questions se sont très vite posées à l'ergonomie en conception: « *Comment aller au-delà des habitudes des utilisateurs ?* » Et « *Comment analyser*

une situation qui n'existe pas encore ? » Pour la première question, nous savons depuis longtemps qu'il ne suffit pas de demander à l'usager ce qu'il veut pour innover, car cela nous conduirait souvent à l'immobilisme. Henry Ford disait déjà : « *Si j'avais demandé aux gens ce qu'ils voulaient, ils m'auraient répondu des chevaux plus rapides* ». En effet, cette objection a longtemps été avancée pour rejeter la CONCEPTION PARTICIPATIVE, l'utilisateur étant décrit comme une force de résistance au changement. Nous savons aujourd'hui que cela est avant tout une question de méthode et l'ergonome joue un rôle central dans son bon déroulement. Il doit aider les opérateurs à se projeter dans les situations futures et traduire leurs contributions en termes utiles aux concepteurs. Il s'agit de mettre en place les moyens permettant une véritable intelligibilité mutuelle et cela constitue, pour les différentes parties, un réel apprentissage (Darses & Reuzeau, 2004).

Méthodes de prototypage

A la deuxième question, « *Comment analyser une situation qui n'existe pas encore ?* », une réponse a été apportée par toute une série de méthodes permettant la création d'artéfact intermédiaire, afin de permettre à l'utilisateur de se projeter à chaque étape du processus de création. En effet, dès le début des années 80, la révolution du développement itératif [19]et la nécessité de gérer tôt les problèmes liés à l'interface utilisateur, ont conduit au développement du PROTOTYPAGE (Tanner & Buxton, 1985). D'ailleurs, une grande partie du travail en IHM, à cette époque, a consisté à créer un environnement et des outils de prototypage de meilleures qualités. Cela a conduit à la séparation de la couche de l'interface utilisateur et de celle de l'application logicielle, afin d'en faciliter la reconception (Carroll, 1997). Dans la prolongation de cette approche, le PROTOTYPAGE RAPIDE s'est vite imposé en tant que standard (Wasserman & Shewmake, 1982). Son but a été de développer des prototypes en une fraction du temps qu'il aurait fallu pour développer un système opérant. En raccourcissant le cycle prototype-évaluation, l'équipe de conception peut évaluer plusieurs alternatives et itérer plusieurs fois, augmentant ainsi la probabilité de trouver une solution qui répond avec succès aux besoins de l'utilisateur. Ils servent également à éliminer les alternatives non prometteuses sans que cela ne coûte trop cher en temps ou en développement (Beaudouin-Lafon & Mackay, 2009). L'important, à cette étape, est de ne pas figer trop tôt la solution qui sera choisie. Pour éviter les écueils, cette phase est généralement réalisée en deux grandes étapes : le MAQUETTAGE DE BAS NIVEAU (« *LOW FIDELITY PROTOTYPING* »), c'est à dire le maquettage des écrans sans habillage graphique, et le PROTOTYPAGE DE HAUT NIVEAU (« *HIGH FIDELITY PROTOTYPING* »), qui présente exactement le comportement et la présentation finale des écrans du futur outil. Nielsen

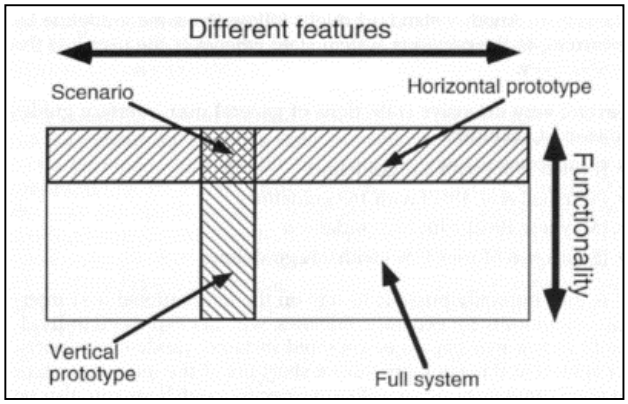


Figure 22 – Les deux dimensions du prototypage *(Nielsen, 1993)*

[19]Le développement itératif s'organise en une série de développement court, de durée fixe, nommée itérations. On parle également de développement incrémental car chaque développement par itération enrichit l'existant. C'est le contraire du développement en cascade, où chaque étape n'est réalisé qu'une seul fois.

(1993) parle également de PROTOTYPAGE « HORIZONTAL » et « VERTICAL » (Figure 22). Le PROTOTYPAGE HORIZONTAL correspond à un maquettage statique d'une interface permettant de rendre compte de l'agencement général de l'interface et des éléments qui la composent. Le PROTOTYPAGE VERTICAL permet, par une succession d'écrans, de dérouler un scénario d'utilisation typique ou de simuler une tâche complète et significative du produit.

Néanmoins, le panel des techniques est beaucoup plus vaste que ces catégorisations élémentaires et se décline du plus grossier au plus proche du rendu final :

- **Les SCENARIOS D'UTILISATION (ou STORYBOARDING)** : Un STORY-BOARD est un ensemble de dessins décrivant ce que le produit fait et comment il est utilisé pour accomplir un ensemble de tâches dans un scénario réel d'utilisation (Figure 23). Cette technique permet d'avoir un retour utilisateur sur une idée d'usage sans avoir à développer de prototype, ni même d'interface (Landay & Myers, 1996; Rosson & Carroll, 2001). C'est pour Nielsen la version la plus minimale du prototype (Nielsen, 1993) car elle combine les limites du PROTOTYPE HORIZONTAL (les utilisateurs n'ont pas accès aux informations de l'interface) et du PROTOTYPE VERTICAL (les utilisateurs ne peuvent pas se déplacer librement au travers du système).
- **Le PROTOTYPAGE PAPIER-CRAYON** : C'est une forme très rapide de prototypage qui peut se réaliser à partir d'un crayon, de papier ou de transparents pour figurer grossièrement les aspects d'un système interactif (par exemple, Muller, 1991). En jouant le rôle de l'utilisateur et du système, les concepteurs peuvent se faire une idée rapide d'une grande variété d'interactions et de l'organisation des interfaces dans une période de temps très courte (Rettig, 1994). Une variante informatisée a également vu le jour par la suite, le WIREFRAMING (ou ZONNING) qui se réalise avec des outils simples d'utilisation tel que Balsamiq[20] (Figure 24).
- **Le WIZARD OF OZ :** Il est utile parfois de donner aux utilisateurs l'impression de travailler avec un système avant même qu'il existe. Pour cela, Kelley (1983) a mis au point une technique surnommée MAGICIEN D'OZ , car elle se base sur une scène du film du même nom. Dans le film, un homme se fait passer pour un magicien en utilisant une série de mécanismes cachés pour créer l'illusion de pouvoirs magiques. La méthode du même nom opère ainsi avec le même principe : un utilisateur interagit avec une interface qui semble être liée à un système fonctionnel, mais qui est, en réalité, simulée par un complice caché quelque part. C'est une technique très utile pour tester en amont la valeur d'un ensemble

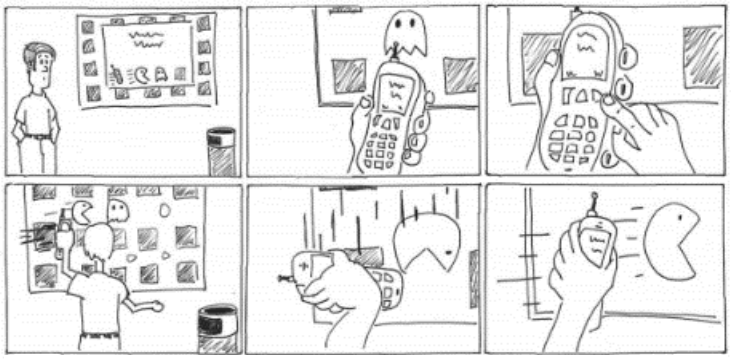


Figure 23 – Exemple de story-board *(Truong, Hayes, & Abowd, 2006)*

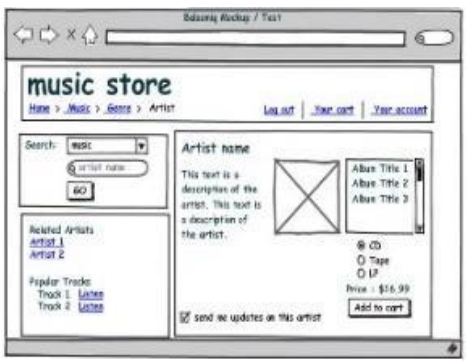


Figure 24 – Exemple de maquette balsamik

[20] http://balsamiq.com

de fonctionnalités, coûteuses ou difficiles à développer (Beaudouin-Lafon & Mackay, 2009).

- **Le PROTOTYPAGE INTERACTIF / HAUTE DEFINITION :** ce dernier type de prototypage permet une forme d'interaction qui se rapproche plus ou moins du logiciel final. Une manière assez simple de simuler la succession d'écrans peut être réalisée en utilisant des applications grand public comme PowerPoint, Adobe Photoshop ou en créant une succession de pages HTML. L'utilisation d'outils spécialisés, tels que *Hypercard* (Goodman, 1987) and *Macromedia Director* a été une solution très courante à la fin des années 80. Néanmoins, depuis une vingtaine d'années, l'évolution des boîtes à outils et des environnements de construction d'interfaces utilisateur, dont les productions peuvent être directement injectées dans l'application finale, a rendu beaucoup plus facile l'exercice de prototypage interactif de nos jours (Beaudouin-Lafon & Mackay, 2009).

Méthodes d'évaluation en cours de conception

Ces différentes méthodes de génération d'artefacts permettent de réaliser des évaluations tout au long du cycle de développement. Deux approches évaluatives sont couramment citées : l'APPROCHE EMPIRIQUE, qui consiste à récupérer directement les performances ou opinions des utilisateurs ; et l'APPROCHE ANALYTIQUE, qui se base sur un ensemble de référents (théories, modèles, …) afin d'évaluer l'artefact et corriger les éventuels problèmes. De même, il existe schématiquement deux familles d'EVALUATIONS EMPIRIQUES : les EVALUATIONS FORMATIVES et SOMMATIVES. L'objectif principal de l'EVALUATION FORMATIVE est de diagnostiquer rapidement et en détail les forces et faiblesses d'un produit, pour savoir comment la conception peut être améliorée. La méthode la plus typique pour l'EVALUATION FORMATIVE est le parcours d'un produit avec la méthode de VERBALISATION A HAUTE VOIX (THINK ALOUD ; Lewis, 1982). En revanche, l'EVALUATION SOMMATIVE vise à évaluer la qualité globale d'un produit, par exemple, pour décider entre des alternatives, ou encore pour analyser le niveau de qualité d'un produit par rapport à sa concurrence. L'EVALUATION SOMMATIVE est souvent conçue comme un test de mesure. L'EVALUATION FORMATIVE est utilisée tout au long du projet alors que l'EVALUATION SOMMATIVE est moins fréquente et plus tardive, car demandant plus d'efforts et une maturation plus avancée du prototype. L'EVALUATION SOMMATIVE est illustrée par les travaux de Robert et Moran (1982, 1983), qui expose bien le changement de philosophie avec les tests précédents. Dans les années 80, l'heure de gloire est aux éditeurs de texte, qui étaient les « *souris blanches des IHM* », car ils pullulaient sur le marché (Green, 1984, cité par Grudin, 2005). Progressivement, l'étude systématique des éditeurs de texte devint impossible, car l'activité se complexifiait. Pour contourner ce problème, Robert et Moran (1982, 1983) proposèrent une évaluation standardisée des éditeurs de texte. Elle s'opposa aux évaluations spécifiques, qui cherchaient à déterminer les bénéfices d'un produit ou d'une fonctionnalité bien précise, dans une situation bien particulière. Elle ne s'attachait pas non plus à couvrir tous les aspects de l'usage des éditeurs de texte. Au contraire, l'évaluation se concentrait sur les propriétés communes des éditeurs de texte et des activités types, afin de pouvoir les comparer sur leur cœur d'activité. L'évaluation chercha à mettre en avant certaines qualités, telles que l'objectivité (la méthodologie ne favorise pas un type d'éditeur de texte en particulier), la complétude (la méthodologie se concentre sur les dimensions de l'usage fondamentales) et la

facilité de mise en œuvre (peu de ressources nécessaires et réalisable par un « *non-psychologue* »). De plus, l'évaluation se concentra sur des indicateurs quantitatifs, comme le temps nécessaire et le coût des erreurs dans la réalisation de certaines tâches d'édition de base par des experts, ou encore la facilité d'apprentissage de ces tâches par des novices. Robert et Moran (1982, 1983) ouvrèrent ainsi une voie méthodologique quantitative différente à l'A/B TESTING classique. Elle permit de mesurer l'utilisabilité d'un logiciel de manière holistique, en se concentrant sur les tâches centrales de l'activité en question. Le TEST D'UTILISABILITE sous cette forme sera repris, amélioré, adapté, documenté (Dumas & Fox, 2009; Molich & Dumas, 2008) et constitue encore de nos jours une pratique vivace dans le domaine de l'évaluation IHM.

À côté des EVALUATIONS FORMATIVES et SOMMATIVES, l'utilisation de QUESTIONNAIRE D'UTILISABILITE se popularisa (Root & Draper, 1983), et permit de compléter les données recueillies par les TESTS D'UTILISABILITE. Toute une série d'initiatives se développa pour créer et valider des questionnaires faciles à administrer, fiables et valides (Tableau 2).

Tableau 2 – Liste des questionnaires d'utilisabilité existants

Année	Nom et acronyme du questionnaire	Référence
1986	Usability Scale (SUS)	Brook (1996)
1988	Questionnaire for User Interface Satisfaction (QUIS)	Chin, Diehl, & Norman (1988)
1989	Perceived Usefulness and Ease of Use (PUEU)	Davis (1989)
1991	After Scenario Questionnaire (ASQ)	Lewis (1991)
1993	Software Usability Measurement Inventory (SUMI)	Kirakowski & Corbett, (1993)
	Nielsen's Attributes of Usability (NAU)	Nielsen (1993)
1995	Computer System Usability Questionnaire (CSUQ)	Lewis (1995)
1997	Purdue Usability Testing Questionnaire (PUTQ)	Lin et al. (1997)
1998	Web Site Analysis and MeasureMent Inventory (WAMMI)	Kirakowski & Cierlik (1998)
2001	USE Questionnaire	Lund (2001)

Se développa également toutes une série de méthodes d'EVALUATION EXPERTES, telles que les EVALUATIONS HEURISTIQUES ou SCRIPTEES (ex : guidelines). À l'inverse des méthodes d'EVALUATION EMPIRIQUES, qui reposent directement sur les performances ou les opinions recueillies, ces méthodes d'EVALUATION ANALYTIQUES se basent sur l'examen d'un produit à partir d'un ensemble de théories ou de modèles existants (Bastien & Scapin, 1993). Elles ont l'avantage de pouvoir être pratiquées tôt dans le cycle de conception, là où l'emploi de TESTS D'UTILISABILITE est impossible. De plus, elles peuvent être utilisées quand les ressources disponibles (argent, temps, évaluateurs entraînés) font défaut (Nielsen, 1989). Cependant, elles n'ont pas l'intention de remplacer les TESTS D'UTILISABILITE classiques, mais d'équilibrer le répertoire de techniques utilisées (Jeffries & Desurvire, 1992). L'EVALUATION HEURISTIQUE consiste à inspecter une interface en évaluant le degré d'application d'une liste de lignes directrices, établies de façon à détecter les aspects positifs et négatifs du point de vue de l'utilisabilité, afin d'en déduire des améliorations. Elle est réalisée par plusieurs évaluateurs experts. Les « *Dix heuristiques d'utilisabilité* » de Nielsen (1994) sont les plus utilisés, et ont été conçues pour évaluer les logiciels et les sites Web. Les CRITERES ERGONOMIQUES de Bastien et Scapin font également partie de cette famille de méthodes d'évaluation (Bastien & Scapin, 1993). Ces derniers font référence dans le domaine grâce à la qualité méthodologique de leur développement. En effet, ces derniers ont été formalisés suite à la synthèse de plus de 800

recommandations, provenant de la recherche ergonomique, et ont été validés expérimentalement. De plus, ces critères, destinés à la base à l'ergonomie des logiciels, ont été également étendus au web (Bastien, Leulier, & Scapin, 1998) et aux interactions homme-environnement virtuel (Bach & Scapin, 2003). L'autre famille de METHODES ANALYTIQUES développées à l'époque est la méthode dite de BALADE COGNITIVE (COGNITIVE WALKTHROUGH). Dans cette méthode, un petit nombre d'experts se « *promènent* » à travers les écrans de l'application en simulant la réalisation de la tâche par un utilisateur (Bias, 1991; Polson, Lewis, Rieman, & Wharton, 1992; Wharton, Rieman, Lewis, & Polson, 1994). Pendant l'inspection, les évaluateurs examinent toutes les actions que l'utilisateur est censé remplir. Pour chacune de ces actions, ils doivent imaginer ce que l'utilisateur serait tenté de faire, sur la base de savoirs supposés et des objectifs de l'utilisateur. Ensuite, les évaluateurs doivent comparer les actions hypothétiques de l'utilisateur en fonction de ce que le système permet afin d'en relever les problèmes probables.

Ce courant de réduction des coûts d'évaluation est porté par Nielsen dès 1989, sous le nom d'INGENIERIE D'UTILISABILITE « *DISCOUNT* » (Nielsen, 1989). Il revendiquera l'utilisation de test utilisateur formatif de cinq participants, le prototypage papier et l'évaluation heuristique. Il déclarera que cet IHM de « *Guérilla* » (Nielsen, 1994a) , représenté par l'utilisabilité « *Discount* » est préférable à l'utilisabilité de « *Luxe* » car elle permet de fournir aux concepteurs des données sur l'utilisabilité de façon fréquente via un cycle d'itération rapide et précoce (Nielsen, 1989).

Les méthodes d'évaluation face à l'avènement de l'informatique sociale

Dans les années 80, les chercheurs se retrouvèrent confrontés à un nombre grandissant de technologies informatiques permettant d'assister la communication ou la collaboration entre individus. La première technologie à attirer l'attention des chercheurs est l'e-mail, qui est rapidement suivi par les outils de conférence textuelle, auditive puis vidéo, les outils d'écriture collaborative, la messagerie instantanée, suivie par la mise en place d'Ethernet, puis de l'internet. L'archétype de l'utilisateur solitaire, recherchant et traitant de l'information seul sur son ordinateur s'estompa rapidement, pour être remplacé par l'archétype d'un collectif de personnes travaillant ensemble, à partir de différents endroits et à différents moments. (Carroll, 1997). L'impact sur les modèles d'évaluation fut immédiat, en agrandissant le nombre de paramètres pouvant jouer sur la qualité d'une application (Figure 25). Cela provoqua l'intégration d'un plus grand nombre de perspectives : utilisabilité, psychologie individuelle, dynamique de groupe, efficience de la communication, impacts sur les structures organisationnelles, culturelles et sociétales, … (Ross, Ramage, & Rogers, 1995). Ainsi, au milieu des années 80, trois communautés de recherche se regroupèrent pour mettre leur effort en commun autour de ce nouveau domaine émergent : (i) des ingénieurs, travaillant sur

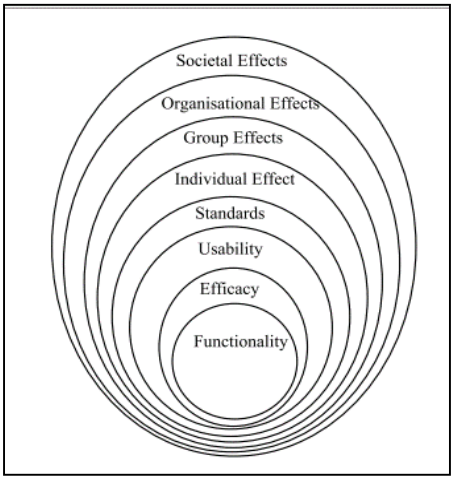


Figure 25 – Les 8 niveaux d'évaluation de Ramage (1999)

l'informatisation et la mise en place des systèmes d'information d'entreprise ; (ii) des chercheurs venus du management et travaillant sur les systèmes de décisions de groupe (« *GDSS* »), les systèmes experts et les EIS (« *Executive Information System* ») ; (iii) des chercheurs en IHM, qui, avec la prolifération des réseaux locaux, sont à la quête de la « *Killer Apps* », capable de soutenir le travail des équipes (Grudin, 2012a). Ils créèrent ensemble le domaine du « *Computer Supported Cooperative Work* » (« *CSCW* »), dont la première conférence eut lieu en 1986 et regroupa de nombreux psychologues, sociologues, anthropologues, ingénieurs logiciels et manager des systèmes d'information (Figure 26).

De manière intéressante, ce regroupement connut en son sein les mêmes problèmes d'éclatement du domaine que pour les IHM et les facteurs humains quelques années plus tôt. Peu de temps après la création du domaine, les chercheurs issus de la communauté CSCW Nord-Américaine et ceux de la branche européenne, soulevèrent un certain nombre de différences entre eux, concernant tant leurs intérêts respectifs que leurs méthodes de recherche. Depuis lors, les Européens organisèrent leurs propres conférences CSCW, dont la première eut lieu en 1989. De même, de nombreux chercheurs initialement venus du champ du management des systèmes d'information américains retournèrent très rapidement dans leur domaine académique d'origine, l'IS (« *Information System* »), à forte coloration managériale et organisationnelle. L'approche de la communauté de recherche européenne a toujours été attachée au contexte du travail en entreprise, un peu comme les facteurs humains à la même époque : elle prend le W (« *Work* ») de l'acronyme CSCW très au sérieux (Grudin, 2012b). De ce fait, leurs enjeux ont été de développer des solutions collaboratives pour des systèmes complexes tels que l'on les retrouve dans les domaines de la santé, des industries manufacturières ou des télécommunications. Leurs pratiques se détournent de la construction théorique et de l' « *Hypothesis Testing* » car ces méthodes sont peu adaptées pour tenir compte des spécificités des organisations de travail complexes et particulières. Ces derniers se tournent donc plus naturellement vers une analyse du travail classique et qualitative rendant compte plus spécifiquement des nécessités de chaque organisation. Les chercheurs Nord-Américains du domaine des IS s'intéressent également au monde du travail, mais sous l'angle du management

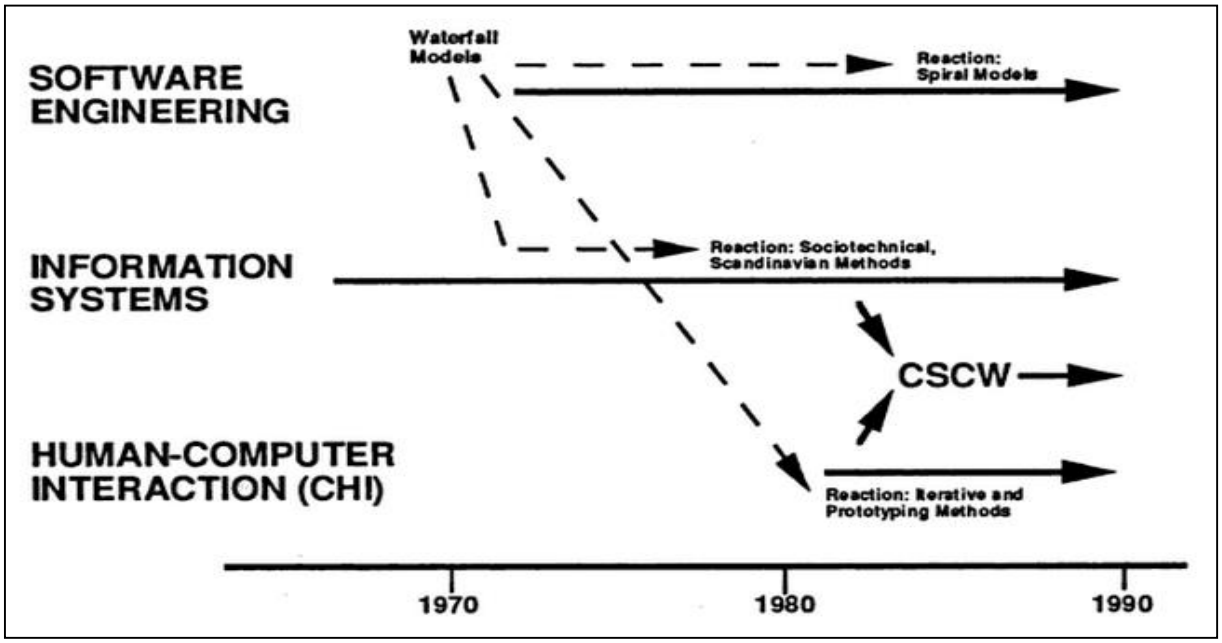


Figure 26 – Evolution des domaines académiques et naissance du champs des CSCW *(Grudin & Poltrock, 1995)*

et des stratégies organisationnelles. Là où les Européens privilégient l'utilisation de méthodes d'évaluation au contact des opérateurs, démêlant le travail « *prescrit* » du travail « *réel* », les chercheurs en IS optent plutôt pour une évaluation hypothético-déductive et statistique. En effet, le domaine des IS s'appuit fortement sur la construction théorique et leur utilisation dans la recherche est en augmentation (Pettigrew & McKechnie, 2001). Les apports théoriques proviennent à 45% des sciences sociales, 20% des sciences « *dures* », 5% des humanités et le reste (30%) du domaine des IS lui-même (Pettigrew & McKechnie, 2001). Parmi ces théories, plus de 10% proviennent du monde de l'économie (K. R. Larsen, Allen, Vance, & Eargle, 2014), ce qui confirme une vision de la gestion de l'entreprise à l' « *américaine* », c'est-à-dire poussée par le « *top management* ».

L'approche de la communauté CSCW Nord-américaine, quant à elle, s'est tournée très rapidement vers les usagers non professionnels. Leur méthode d'évaluation s'est donc inspirée fortement de celles des IHMs, c'est-à-dire privilégiant le prototypage, les tests d'utilisabilité, et un intérêt moindre pour la construction théorique. Deux articles, sortis à l'occasion des 20 ans des CSCW en 2006, nous éclairent plus précisément sur les intérêts et les méthodes du domaine depuis sa création. Jacovi, Soroka, Gilboa-Freedman, Ur, Shahar, et Marmasse (2006) ont analysé le graphe des citations des 465 papiers pour identifier les grands clusters dans le domaine. Ils en ont identifié huit, dont les deux les plus importants correspondent à une partie « *sciences sociales* » et à une partie « informatique » :

- Le cluster des sciences sociales comprend des articles sur les théories et les modèles du domaine, sur la méthode ethnographique et les études utilisateurs (83 articles) ;
- Le cluster informatique comprend des articles sur l'architecture logicielle ou sur des sujets techniques (82 documents);
- Le troisième plus grand groupe (43 articles) comprend les outils de décision et de réunions, les espaces multimédias partagés, et les outils de conférences ;
- Un quatrième groupe est composé de 12 articles sur la messagerie instantanée, les espaces sociaux, et de la notion de présence ;
- Le cinquième groupe est composé de sept articles sur l'utilisation des outils informatiques tels que le courrier électronique en milieu de travail ;
- Les grappes restantes (cinq articles chacun) sont centrées sur la conception des « *Groupwares* » et sur la conscience de l'espace de travail (« *Workspace Awareness* ») ; la gestion informatique et des systèmes d'information; et la communication vidéo et les espaces visuels partagés ;

Tableau 3 - Matrice Espace / Temps de Johansen (1988)

	Même temps (synchrone)	Temps différents (asynchrone)
Même lieu (colocalisation)	Salle de décision, table partagé, affichage mural, …	Salle d'équipe, *Groupware* de travail différé, gestion de projet
Lieux différents (à distance)	Conférence vidéo, messagerie instantané, chat, MUD, mondes virtuels, écran partagés, …	Email, forum, blogs, outils de conférence asynchrone, calendrier de groupes, wikis, …

En complément, le travail de Convertino, Kannampallil et Councill (2006) nous montre que le poids des théories dans la communauté CSCW Nord-américaine est assez faible. En effet,

l'étude nous informe que pour les trois premières conférences, 30% des papiers s'appuyaient sur une contribution théorique, mais que ce chiffre a rapidement baissé sous la barre des 10%, dès que les chercheurs en IS désertèrent le champ.

Nous avons vu dans le chapitre sur le deuxième paradigme de recherche que les psychologues cognitifs des IHMs s'étaient attelés initialement à la construction d'un socle théorique et scientifique pour guider la conception. À la fin des années 1980, le rythme de l'évolution technologique avait calmé ces ambitions. À ses débuts, le domaine des CSCW tenta également l'expérience, mais durant quelques années seulement. De plus, la plupart des invocations théoriques qui en résulta, se contenta d'étendre des théories déjà développées ailleurs, ou avec de petits ajustements et des mises à l'épreuve (Grudin, 2012a). En effet, l'accélération du rythme des changements technologiques et des usages associés eut pour effet de créer toutes sortes de difficultés dans la création d'un savoir durable et utile au domaine. Grudin (2012a) cita de nombreux exemples d'étude, qui furent dépassés par la vitesse fulgurante d'évolution du champ. Par exemple, dans une étude rigoureuse de Kraut et al. (1998), ces derniers constatèrent que l'utilisation d'Internet avait un impact négatif sur le développement social ; seulement quelques années plus tard, l'analyse de données subséquentes montra qu'un changement dans l'expérience, la technologie ou internet lui-même, avait entrainé une dissipation de cet effet (Kraut et al., 2002). De manière similaire, le « *paradoxe de productivité* », qui reflétait le fait que les organisations ne réalisaient pas de bénéfice à la hauteur de leur investissement en informatique, fut formellement présenté par Brynjolfsson (1993) pour être réfuté seulement cinq années plus tard (Brynjolfsson & Hitt, 1998).

L'instabilité du savoir a également touché les efforts de taxonomisation des CSCW. Cela constitue pourtant un exercice utile dans tout champ disciplinaire se constituant. Dans le domaine des CSCW, ces outils a-théoriques de classification permettent aux chercheurs de nommer le type d'application sur lesquels portent leurs recherches, et ainsi de pouvoir constituer un savoir expert à son sujet, réutilisable par les pairs. Une des premières à s'imposer est la matrice espace-temps de Johansen (1988) qui classe les *Groupwares* en quatre types : synchrone dans un même lieu, synchrone dans des lieux différents, asynchrone dans un même lieu, et asynchrone dans des lieux différents (tableau 3). Une autre façon de catégoriser les applications est le modèle de coopération 3C (Figure 27), qui se concentre sur les types d'interactions dans le travail de groupe (Ellis, Gibbs, & Rein, 1991). Elle permet de positionner les applications selon que ces dernières favorisent la communication (échange d'information), la coordination (gestion des personnes, des activités ou des ressources) ou la coopération (opérations conjointes). Des taxonomies hybridant ces deux types de catégorisations virent le jour peu de temps après, comme celle proposée par Poltrock & Grudin

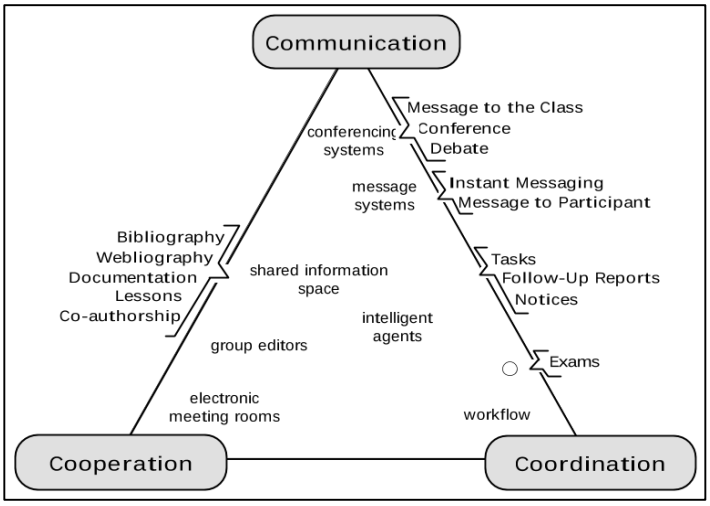


Figure 27 – Modèle 3C, adapté de Reimer et al. *(2009)*

(1998), à deux dimensions : la temporalité (synchrone / asynchrone) et le mode de collaboration (communication, partage d'informations et coordination). De même, Penichet et al. (2007) proposa un modèle de classification qui combine les modes d'interactions de groupe et les caractéristiques spatio-temporelles, soit 63 combinaisons valides. Des modèles taxonomiques plus récents, comme celui de Cruz et al. (2012), vont jusqu'à présenter plus d'une quinzaine de dimensions issues de l'état de l'art (modèle 3C, classification d'espace/temps, conscience de l'environnement de travail, mécanismes de régulations, dynamiques de groupes, …). Ainsi, si l'effort de classification se poursuit encore, ses objectifs paraissent maintenant obsolètes tant le domaine se diversifie et se complexifie rapidement (Penichet et al., 2007). C'est d'autant plus vrai avec la tendance des mashups qui permet une recombinaison presque infinie des différents types d'informations et des médiums de communication (Jackson, 2009; Verdot, Saidi, & Fournigault, 2011). La diversité des applications brise le cloisonnement taxonomique précédemment établi et il devient difficile de créer un savoir assez général pour être appliqué à des situations de plus en plus particulières. L'accent fut donc mis sur des méthodes permettant d'explorer et de valider des artéfacts ayant chacun leur particularité, telles que l'approche de la THEORIE ANCREE (« *GROUNDED THEORY* »), visant par une démarche qualitative à former des connaissances, non pas à partir d'hypothèses prédéterminées, mais à partir des données et des situations de terrain (Annells, 2011). En effet, à cette époque, de nombreux ethnologues rejoignirent les rangs des chercheurs dans les deux branches de la communauté CSCW, Américaine et Européenne (Grudin & Poltrock, 2012). Un état de l'art d'Antunes et al. (2012) des méthodes d'évaluation pour les systèmes collaboratifs (Tableau 4), montrent bien l'évolution des tendances dans le domaine :

- **1er temps : l'adaptation des méthodes d'évaluation mono-utilisateur, développée dans le domaine de l'interaction homme-machine, dans le contexte spécifique des systèmes de collaboration**. Cela s'est produit, par exemple, avec les méthodes de BALADE COGNITIVE (WALKTHROUGHS STRUCTURES, WALKTHROUGHS COGNITIFS, WALKTHROUGHS POUR GROUPWARE), l'EVALUATION HEURISTIQUE (EVALUATION HEURISTIQUE, EVALUATION HEURISTIQUE BASEE SUR LES MECANISMES DE COLLABORATION) et l'évaluation basée sur les scénarios (STORYBOARDING).
- **2ème temps : l'assimilation de perspectives et de techniques d'autres domaines, au-delà des méthodes classiques de conception technologique.** L'exemple le plus évident est l'ethnographie (ETUDES D'OBSERVATION, ETHNOGRAPHIE « *QUICK AND DIRTY* », ANALYSE DU TRAVAIL), mais les sciences cognitives semblent également avoir un impact (KLM, WALKTHROUGHS COGNITIFS, GOMS).
- **3ème temps : la complexification croissante du modèle d'évaluation.** En effet, la plupart des premières méthodes utilisées (ex : WALKTHROUGHS STRUCTURES, APPROCHE KLM, METHODES D'UTILISABILITE « DISCOUNTS ») paraissent se concentrer sur des variables très spécifiques, mesurées dans des conditions contrôlées, tandis que les méthodes plus récentes semblent considérer des facteurs contextuels plus larges (EVALUATION MULTIFACETTE, EVALUATION DE LA COLLABORATION CO-LOCALISEE, APPROCHE BASEE SUR LE CYCLE DE VIE).

De plus, d'autres études du domaine semblent indiquer qu'un tiers des systèmes collaboratifs ne sont pas évalués de manière formelle (Pinelle & Gutwin, 2000), et, quand bien même cela serait le cas, elles sont couramment réalisées en fonction des intérêts des évaluateurs et de

l'adéquation pratique du contexte applicatif (Greenberg & Buxton, 2008; Inkpen, Mandryk, Dimicco, & Scott, 2004).

Tableau 4 - Liste chronologique des méthodes d'évaluation dans le domaine des systèmes collaboratifs

Année	Auteurs	Méthode d'évaluation
1978	Yourdon (1978)	STRUCTURED WALKTHROUGHS
1980	Card et al. (1980b)	KEYSTROKE-LEVEL MODEL (KLM)
1987	Suchman (1987)	ETHNOMETHODOLOGICAL STUDIES
1989	Nielsen (1989)	DISCOUNT USABILITY ENGINEERING
1990	Nielsen & Molich (1990)	HEURISTIC EVALUATION
1991	Tang (1991)	OBSERVATIONAL STUDIES
	Bias (1991)	INTERFACE WALKTHROUGHS
1992	Polson et al. (1992)	COGNITIVE WALKTHROUGHS
	Rowley & Rhoades (1992)	COGNITIVE JOGTHROUGH
1993	Urquijo et al. (1993)	BREAKDOWN ANALYSIS
1994	Wharton et al. (1994)	COGNITIVE WALKTHROUGHS
	Twidale et al. (1994)	SITUATED EVALUATION
	Nielsen (1994b)	USABILITY INSPECTION
	Nielsen (1994b)	HEURISTIC EVALUATION
	Ereback and Höök (1994)	COGNITIVE WALKTHROUGH
	Bias (1994)	PLURALISTIC USABILITY WALKTHROUGH
	Hughes et al. (1994a)	QUICK-AND-DIRTY ETHNOGRAPHY
	Hughes et al. (1994b)	EVALUATIVE ETHNOGRAPHY
1995	Plowman et al. (1995)	WORKPLACE STUDIES
1996	Gutwin et al. (1996)	USABILITY STUDIES
	Van Der Veer et al. (1996)	GROUPWARE TASK ANALYSIS
1997	Baeza-Yates & Pino (1997)	FORMAL EVALUATION OF COLLABORATIVE WORK
1998	Stiemerling & Cremers (1998)	COOPERATION SCENARIOS
	Ruhleder and Jordan (1998)	VIDEO-BASED INTERACTION ANALYSIS
	Briggs et al. (1998)	TECHNOLOGY TRANSITION MODEL
1999	Neale and Carroll (1999)	MULTI-FACETED EVALUATION FOR COMPLEX, DISTRIBUTED ACTIVITIES
	Gutwin & Greenberg (1999)	EVALUATION OF WORKSPACE AWARENESS
2000	Gutwin & Greenberg (2000)	MECHANICS OF COLLABORATION
	Carroll (2000)	SCENARIO-BASED DESIGN
	Van Der Veer (2000)	TASK-BASED GROUPWARE DESIGN
2001	Steves et al. (2001)	USAGE EVALUATION
	Baker et al. (2001)	HEURISTIC EVALUATION BASED ON THE MECHANICS OF COLLABORATION
	Sonnenwald et al. (2001)	INNOVATION DIFFUSION THEORY
2002	Baker et al. (2002)	GROUPWARE HEURISTIC EVALUATION
	Cockton & Woolrych (2002)	DISCOUNT METHODS
	Pinelle & Gutwin (2002)	GROUPWARE WALKTHROUGH
2003	Pinelle et al. (2003)	COLLABORATION USABILITY ANALYSIS
	Antunes and Costa (2003)	PERCEIVED VALUE
2004	Haynes et al. (2004)	SCENARIO-BASED EVALUATION
	Convertino et al. (2004)	ACTIVITY AWARENESS
	Humphries et al. (2004)	LABORATORY SIMULATION METHODS
	Inkpen et al. (2004)	EVALUATING COLLABORATION IN CO-LOCATED ENVIRONMENTS
	Kieras and Santoro (2004)	COMPUTATIONAL GOMS
	Briggs et al. (2004)	SATISFACTION ATTAINMENT THEORY
2005	Vizcaíno et al. (2005)	KNOWLEDGE MANAGEMENT APPROACH
2006	Baeza-Yates and Pino (2006)	PERFORMANCE ANALYSIS
	Antunes et al. (2006)	HUMAN PERFORMANCE MODELS
2008	Pinelle and Gutwin (2008)	TABLETOP COLLABORATION USABILITY ANALYSIS

Dernièrement, même les conférences CSCW et ECSCW continuent d'exister, le domaine s'est un peu dissout dans le domaine plus large des applications numériques connectées. En effet, avec l'avènement d'Internet, rares sont les applications qui ne possèdent pas de mécanisme de coordination, de collaboration ou d'échange d'information. On constat que le livre d'Ackerman et al. (2008) est l'un des derniers qui tente de théoriser ce domaine en gardant l'acronyme CSCW dans son titre. D'ailleurs, l'analyse moderne de cet acronyme par Grudin et Poltrock (2012) montre que chacun de ces termes ont perdu de leur pertinence : Le C (« *Computer* ») était au centre des années 80 mais les technologies numériques sont dorénavant incorporés dans un grand nombre d'appareils numériques bien plus diversifiés ; le S (« *Supported* ») renvoie à une époque où l'informatique était là pour assister une activité. Or aujourd'hui, l'informatique a un rôle central de diffusion de l'information digitale et non plus de support ; le C (« *Coopérative* ») est dépassé par la diversité des rôles que la technologie a explorés dans le cadre des interactions de groupe ; le W (« *Work* ») renvoie à une vision passéiste de l'informatique, car aujourd'hui le domaine a envahi toutes les activités humaines, en passant par les loisirs. Ce changement est également reflété en 2010 par l'éditeur scientifique Springer qui change la signification de l'acronyme de sa série de livre sur les CSCW par « *Collaboration, Sociality, Computation, and the Web* »

Aujourd'hui, le domaine des IHM, dépasse également celui de la simple conception d'interfaces d'entrées/sorties compatibles avec le fonctionnement humain. Ses ambitions sont devenues progressivement plus larges. Grâce à Internet et aux technologies mobiles, il permet de connecter les personnes ensemble, de gérer l'interaction des individus par l'intermédiaire de la technologie : on parle de « *Human-to-Computer-to-Human-Interaction* », ou HCHI (Clubb, 2007). Le domaine prend également du recul, et se centre précisément sur le cœur de l'activité informatique, l'information. L'interfaçage dépasse ainsi le simple rôle de canal de communication entre la machine et l'humain. En conséquence, de nombreux chercheurs anglo-saxons revendiquent le retour à l'utilisation du terme « *informatics* » en place et lieu de « computer science » (Adriaans & van Benthem, 2008b). Lucas (2000) suggère même qu'il faut passer d'un monde centré sur l'informatique (« C*omputer-Centric World* »), à un monde centré sur l'information (« *Information-Centric World* »), c'est-à-dire indépendamment du medium permettant l'échange d'informations (Gershon, 1995). Ce passage d'ergonome à « *Infonome* » (« *From Ergonauts to Infonauts* » ; Brookhuis, 2008) est d'autant plus nécessaire que la convergence numérique rapproche des mediums autrefois distants. Le rapprochement de domaines académiques, autrefois disjoints, engendre des dialogues fertiles et des hybridations. De nouveaux domaines naissent, tels que celui de documentation numérique (Marchionini, 2008), de la télévision interactive (Jain, Evans, & Vinayagamoorthy, 2013) ou de l'e-learning. Pour Cronin (1995, p. 56), les concepts autour de l'information *« fonctionneront comme des aimants ou des attracteurs, attirant hors des disciplines tout contenu la concernant et se restructurera dans un cadre scientifique autour de l'information* ».

En résumé, ce mouvement d'élargissement de la sphère informatique, ainsi que de l'accélération du développement logiciel réorienta les besoins en matière d'évaluation. Ces dernières intervinrent plus tôt dans la conception et s'effectuèrent plus souvent. Elles gagnèrent en variété, en rapidité et en souplesse, même si la demande de réactivité face aux attentes du marché les amena à perdre en précision. De ce fait, les méthodes qualitatives prirent leur envol

face aux méthodes quantitatives. À la fin des années 80, une mutation bien plus profonde du domaine conduira le champ à un mouvement de convergence bien plus global et plus rapide encore. Cela impacta directement les méthodes dont la tendance vers le qualitatif s'amplifia encore. Nous verrons toutefois que l'approche quantitative refera surface, grâce à la conjecture technologique qui va rendre ce retour possible et grandement profitable pour les concepteurs.

Paradigme 3 – L'expérience Utilisateur

À la fin des années 80, le monde de la conception et de l'évaluation Web se retrouva bouleversé par l'effet même de son propre déploiement. En effet, l'efficacité des méthodes et des outils de production logicielle alla grandement accélérer son propre développement. De plus, les incessantes avancées technologiques allèrent lui permettre d'investir bien d'autres domaines que celui de la bureautique. Cela a, à son tour, entraîné l'émergence de nouveaux modes de consommation, une mutation des modèles économiques existants et un jaillissement de valeurs et de demandes nouvelles que le numérique dut satisfaire. Tous ces changements allèrent complexifier les modèles d'évaluations existants et entraîner l'apparition d'un nouveau paradigme de recherche dans le monde du numérique.

Le contexte d'émergence du nouveau paradigme

De nombreux facteurs précurseurs de ce bouleversement prirent leur source en dehors du champ et allèrent progressivement se rejoindre, à la manière de petits ruisseaux qui firent une grande rivière.

L'impact du « bootstrapping » sur la diversification de l'économie numérique

Il a fallu attendre en quelque sorte une quarantaine d'années pour que la vision d'Engelbart se réalise enfin. Ce dernier fit le constat dans les années 50 que l'Homme n'était pas armé pour les défis qui allait l'attendre. Pour lui, seule une accélération du développement de l'intellect humain de manière exponentielle pouvait résoudre ce problème. Il fit l'hypothèse qu'en utilisant l'informatique naissante, il serait capable d'augmenter l'intelligence humaine et ainsi, par coévolution, accélérer leur capacité intellectuelle conjointe. Il dévoua sa vie à cet objectif et testa sa méthode sur sa propre équipe de recherche dès 1962. Cette méthode réflexive du troisième ordre, le « *Bootstrapping* » (Bardini, 2000), eut pour objectif de faire progresser les entreprises intellectuelles selon trois stratégies : (i) résoudre les problèmes, (ii) améliorer la manière de les résoudre, (iii) améliorer la manière d'améliorer. Cette philosophie du développement, utilisée instinctivement et massivement en informatique, conduisit, par effet boule de neige, à une véritable accélération auto-entretenue du développement du monde numérique.

Elle fut permise par la création de processus, méthodes et outils, guidant et automatisant grandement la conception des nouveaux produits numériques. En effet, la mise au point de nouvelles méthodes de programmation et de prototypage évolutif réduisit leurs coûts et temps de fabrication (Beaudouin-Lafon & Mackay, 2009). De plus, les méthodes d'évaluation

« *discounts* », et l'organisation du développement itératif, « *agile* »[21] (Ambler, 2002) ou « *extrême* »[22] (Beck, 1999) participèrent également à ce mouvement. La course à la digitalisation, grâce aux développements technologiques, étendit l'informatique à d'autres domaines qu'elle optimisa à leur tour. Les entreprises se retrouvèrent ainsi obligées de suivre la course à l'armement numérique pour rester compétitives. Ainsi, Microsoft fit fortune grâce à son système d'exploitation Windows, puis grâce à sa suite bureautique Office, dans les années 90. La puissance économique passa du domaine matériel (qui s'écroula, comme IBM) aux puissances logicielles. Le software l'emporta sur le hardware. En effet, la grande flexibilité apportée par la programmation, couplée à un coût et à une vitesse de développement moindre, en fit une arme puissante, facile à produire et facile à distribuer. De nombreux acteurs s'investirent alors dans ce nouvel eldorado, ce qui ne tarda pas à rendre le domaine hyper-compétitif. En effet, les industries passèrent d'un rythme lent, composé d'un petit ensemble d'oligopoles stables, à un environnement caractérisé par une concurrence intense et agressive, frappant rapidement et de manière non conventionnelle (D'Aveni, 1997). Internet accentua encore cette tendance, car il permit l'accession pour les entreprises de marchés très éloignés et donc de mettre en place une compétition de plus en plus globale et mondialisée. La disposition de nombreuses entreprises nouvelles à investir dans d'autres champs que celui, saturé, de la bureautique, va les pousser à s'aventurer vers des domaines tels que l'industrie des loisirs, l'éducation, la communication ou la santé. De ce fait, des champs académiques, encore extérieurs aux IHM, vont progressivement s'y intégrer afin de nourrir de préoccupations inédites la conception et l'évaluation de ces nouvelles applications (Figure 28).

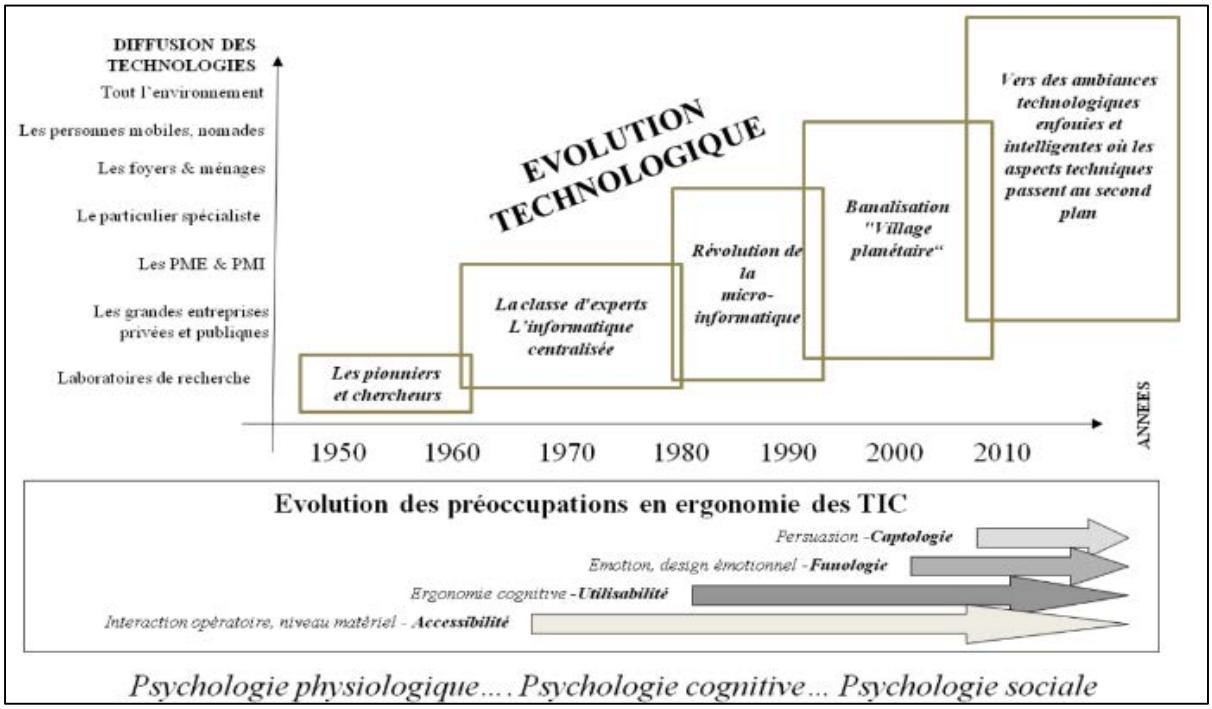


Figure 28 - L'évolution corrélative des technologies, de leur diffusion dans la société et des orientations en psychologie ergonomique. (Brangier & Bastien, 2010)

[21] Se dit d'un processus de développement plus léger, permettant de gagner en réactivité

[22] Méthode de développement informatique où l'implémentation du code se fait par petit incrément et itération, et où des livrables réguliers sont présentés aux clients après chaque cycle de développement

L'industrie des loisirs, la psychologie positive et la « funologie »

À partir des années 80, puis à partir de grands consortiums dans les années 90, le développement des jeux vidéo va grandir de façon exponentielle. L'idée de prendre du plaisir et de s'amuser avec les machines va commencer à s'ancrer dans les esprits. A cette occasion, l'ergonomie redécouvrira que les émotions sont de véritables leviers de l'interaction et que les règles ergonomiques, définies pour le monde professionnel, sont insuffisantes pour assurer l'adéquation et l'adoption des applications dans le domaine des loisirs et de la consommation (Brangier & Bastien, 2010). En effet, pendant des décennies, la communauté IHM considéra qu'un bon système était un système sans problème. C'est-à-dire que l'absence d'éléments négatifs (erreurs, frustrations) fut perçue comme la seule condition nécessaire pour une utilisation de qualité. Or ce sont la présence d'éléments positifs (développement personnel, plaisir) qui fournissent la motivation nécessaire à l'usage du produit (Robert, 2008). Cette approche, tronquée, héritée de l'ère cognitiviste, est pourtant démentie par des travaux antérieurs en psychologie… ce que des chercheurs vont remettre au goût du jour.

Les travaux sur la satisfaction au travail du psychologue américain Herzberg peuvent être pris en exemple. Il montra, avec sa théorie des deux facteurs (le « *Motivator-Hygiene model* » ; Herzberg, Mausner, & Snyderman, 1959) que la satisfaction au travail et l'insatisfaction au travail agissent de manière indépendante. Ainsi, le contraire de la satisfaction n'est pas l'insatisfaction mais l'absence de satisfaction. De même, le contraire de l'insatisfaction est l'absence d'insatisfaction. De ce fait, il distingua deux séries de facteurs. Les premiers sont externes au travail lui-même, telles que la rémunération, les relations avec l'encadrement ou les conditions de travail, que Herzberg appelle « *facteurs d'hygiène* ». Leurs absences ou dysfonctionnements insatisfont les salariés, ce qui les pousse à les réclamer. Mais leurs présences apaisent sans vraiment stimuler. Les seconds facteurs sont plus internes du travail, tels que l'intérêt du travail en lui-même, la reconnaissance, les responsabilités ou le progrès personnel, que Herzberg qualifia de « *facteurs de motivation* ». Ces derniers dynamisent et stimulent la production : ce sont eux qui donnent du cœur à l'ouvrage. Ainsi, en transposant le modèle des deux facteurs d'Herzberg aux IHM, il est facile de voir l'utilisabilité comme un facteur d'hygiène plutôt qu'un facteur de motivation (Schaffer, 2009). Les facteurs motivationnels sont donc ceux qu'il fallait développer pour compléter les modèles IHM classiques.

Ces derniers commencent pourtant à être étudiés scientifiquement, notamment dans une discipline encore peu connue : celle de la psychologie positive. Pourquoi « *positive* » ? Parce que la psychologie, jusqu'il y a peu, s'occupait en majorité des problèmes cliniques, scolaires, ou de productivité au travail. En effet, avant la Seconde Guerre mondiale, la psychologie avait trois missions fondamentales : soigner les maladies mentales, rendre la vie des gens plus productives, et identifier et nourrir les hauts talents. Un psychologue hongrois, Mihály Csíkszentmihály, investira dans les années 70 un champ de recherche complètement différent de ceux généralement étudiés en psychologie, suite à son expérience traumatisante de la guerre. Il déclara : « *En tant qu'enfant, j'ai vu pendant la guerre que quelque chose n'allait pas avec la manière dont les adultes –les grandes personnes en qui j'avais confiance– organisaient leur raisonnement. J'ai donc essayé de trouver un meilleur système pour organiser ma vie* »

(Seligman & Csikszentmihalyi, 2000). Il développa ainsi une discipline nouvelle, s'attachant à l'étude scientifique des forces du fonctionnement optimal et des déterminants du bien-être. Ce mouvement de la psychologie positive fut officiellement lancé, en 1998, par un discours de Martin Seligman, alors président de l'American Psychological Association (Seligman, 1999). Il déclara que la psychologie avait consacré trop d'efforts sur la maladie mentale, négligeant l'autre extrémité du spectre, soit le fonctionnement optimal, le sens et le bonheur. Les thèmes abordés y sont incroyablement variés : l'altruisme, l'amitié, l'amour, le bonheur, la confiance, le courage, la créativité, les émotions, l'empathie, l'engagement, l'espoir, l'humour, la motivation, l'optimisme, les sens de l'existence… soit tout ce qui fait que la vie vaut la peine d'être vécue. Ainsi, les années 1980 ont vu une appréciation croissante de la notion de « *flux* » (« *Flow* »), qu'utilisera Csikszentmihalyi comme cadre conceptuel pour expliquer les interactions utilisateurs positives. Le flux peut être décrit comme un état psychologique complexe, traduisant un sentiment d'expérience optimale et caractérisé par l'engagement dans une activité avec une forte implication, concentration, plaisir, et motivation intrinsèque (M. Csikszentmihalyi, 1975). D'autres notions vont être progressivement intégrées à la discipline, comme les besoins sociaux (Jordan, 1999) ou la notion de stimulation (Hassenzahl, 2004).

Un autre courant théorique, la « *Funologie* », va également s'affirmer dans les années 2000 (Blythe, Monk, Overbeeke, & Wrigh, 2003), mettant en avant un certain nombre de qualités hédoniques dans l'interaction. De manière plus spécifique, de nombreuses méthodes d'évaluation seront adaptées pour le domaine porteur des jeux vidéo, telle que la METHODE RITE (Medlock, Wixon, Terrano, Romero, & Fulton, 2002) ou PLAYTEST (Davis, Steury, & Pagulayan, 2005). Une série d'HEURISTIQUES sera également mise au point (Desurvire, Caplan, & Toth, 2004; Federoff, 2002; Malone, 1982). D'autre part, l'attrait grandissant pour la prise en compte des réponses affectives amena les chercheurs à importer du Japon des techniques d'ingénierie émotionnelle, tel que le KANSEI DESIGN (du japonais « 感性工学 » : « *Kansei Kougaku »*). Son but est le développement ou l'amélioration de produits par la conversion des besoins psychologiques et émotionnels des clients en paramètres de conception (Nagamachi, 2002; Nagamachi, 1995). Du côté informatique, l'exploitation des réponses affectives sera perçue comme une opportunité pour améliorer et enrichir nos moyens d'interagir avec la technologie : ce sera la naissance du courant de l'informatique affective (« *Affective Computing* » ; Picard, 1997). En effet, par la reconnaissance des émotions humaines, à partir d'expressions faciales, posturales, tonales ou physiologiques, ce mouvement essayera de répondre à des problématiques innovantes, telles que l'optimisation de l'état d'attention de l'utilisateur ou la réduction de sa frustration.

Cette évolution des préoccupations en IHM est reflétée par les contributions successives de l'influent psychologue Donald Norman, passant du cognitivisme (i), à l'utilisabilité (ii) puis au désir (iii). (i) En tant que scientifique cognitiviste, à qui l'on doit la notion « d'ingénierie cognitive » (Norman, 1983b), il commença par déclarer, lors de la première conférence CHI en 1983, que la satisfaction utilisateur se base sur « *la vitesse d'utilisation, la facilité d'apprentissage, les connaissances nécessaires et le nombre d'erreurs* » (Norman, 1983a). (ii) En 1988, son célèbre ouvrage « *la psychologie des objets de la vie quotidienne* » (« *The*

Psychology of Everyday Things »)[23], se concentra davantage sur la conception centrée utilisateur et l'utilisation pragmatique du concept d'utilisabilité (Norman, 1988). L'intérêt d'une telle démarche est, par ailleurs, bien illustré par la couverture du livre (Figure 29). (iii) Seize ans plus tard, il publia « *Design Emotionnel : pourquoi aimons-nous ou détestons-nous les objets qui nous entourent* » (« *Emotional Design: Why We Love (or Hate) Everyday Things* »), soulignant dorénavant le rôle notable des émotions et de l'esthétisme dans notre appréciation aux objets (Norman, 2004). On notera ici que le terme « *Design* » ne se traduit plus par « *conception* », montrant, qu'en France, il ne signifie plus la même chose. L'approche de Norman dans ce livre reposa sur une adaptation du modèle ABC de l'attitude (Rosenberg & Hovland, 1960), dont il renomma les trois dimensions pour être compatible avec le champ de la conception. Il avança que tout objet est perçu sur ces trois niveaux (viscéral, comportemental et réflexif) et que toute bonne conception doit prendre en charge ces trois niveaux. La couverture du livre, en montrant une icône du design industriel, le Presse agrumes « *Juicy Salif* » de Philippe Starck (Figure 30), illustra convenablement la nouvelle thèse de Norman, basé sur le pouvoir de séduction des objets.

Figure 29 − Couverture de « *la conception des objets de la vie quotidienne* »

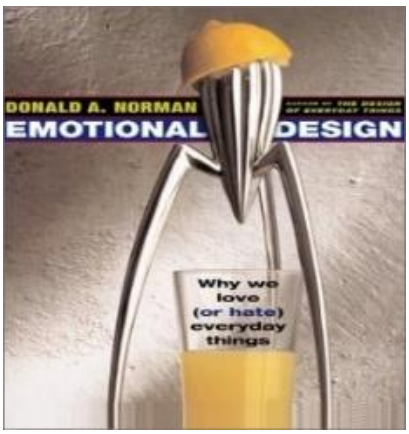


Figure 30 − Couverture de « *Design Emotionnel : pourquoi aimons-nous ou détestons-nous les objets qui nous entourent* »

L'Homo Aestitheticus et le design

Dans les années '50, la société de consommation était un paradis pour les vendeurs. Dans ce monde enchanté, un aspirateur, une machine à laver ou un téléviseur étaient admirés par le simple fait d'exister. Progressivement, l'augmentation du taux d'équipement des ménages et l'assouvissement progressif des besoins primaires par les masses affaiblirent leurs pulsions d'achat. Il fallut donc trouver des moyens de « *ré-enchanter* » la consommation. Au Japon, un moyen fut trouvé via le marketing émotionnel. En effet, durant les années 80, un changement de comportement sur le marché japonais fut observé : le comportement logique et prévisible (en japonais « *Risei* ») cédait progressivement la place à un comportement d'achat plus imprévisible, c'est-à-dire basé sur les sensations/émotions (« *Kensei* », en japonais). En raison de ce changement de comportements, le marché s'adapta et passa d'une production de masse,

[23] Le livre fut renommé un peu plus tard « *la conception des objets de la vie quotidienne* » (« *The Design of Everyday Things* ») pour toucher un public plus large

destinée à tout le monde, à une production individualisée, répondant aux goûts et aux désirs de chaque consommateur (Dentsu, 1985; Fujioka, 1984).

Cette démarche de personnalisation affective s'appuiera, partout dans le monde, sur des stratégies de valorisation assez classique, tel que la création d'une identité produit gratifiante par la publicité, mais également par l'utilisation de l'esthétisme pour étendre le spectre de biens à styliser et donc sujet à être remplacé en fonction de la mode. Ces stratégies ont permis d'intégrer le supplément d'âme qui manquait aux produits jusqu'alors, et de redynamiser les consommations. Dans ce domaine, les Américains furent les champions toute catégorie. L'histoire nous montre que ces derniers ont souvent accompagné leurs forces de productions industrielles par leurs industries culturelles qui les valorisent : *« Trade follows film, rather than the flag* » écrivit un sénateur américain dès 1912 (Frodon, 1998). Ingénieusement, le « *Hard Power* » (militaire) fut remplacé par le « *Soft Power* » (culturel et idéologiques), permettant dans un contexte où la guerre est devenue économique, d'influencer indirectement le comportement des autres par des moyens non coercitifs, souvent même sans qu'il en ait conscience. Ainsi, la créature théorique des économistes néoclassiques, l'*homo œconomicus*, cet être cupide, rationnel froid et socialement atomisé, ressort sous cet angle comme une absurdité. Les individus ont des émotions, des valeurs, une éthique, façonnées par des entités collectives et influencées par les institutions (Stiglitz, Escoto, Chemla, & Chemla, 2010), et qui sont tout à fait exploitables par elles.

Ainsi, les sociétés marchandes exploitent depuis longtemps la part irrationnelle du comportement d'achat. Nous savons aujourd'hui que la désirabilité d'un produit ne tient que peu de son utilité ; elle l'est davantage par les valeurs de séduction qu'elle déploie, c'est-à-dire sociales, émotionnelles, esthétiques. C'est un principe très exploité de nos jours. À ce sujet, Gilles Lipovetsky et Jean Serroy (2013) préciseront que « *partout le réel se construit comme une image en y intégrant une dimension esthétique-émotionnelle devenue centrale dans la compétition que se livrent les marques. Tel est le capitalisme artiste, lequel se caractérise par le poids grandissant des marchés de la sensibilité, par un travail systématique de stylisation des biens et des lieux marchands, par l'intégration généralisée de l'art, du «look» et de l'affect dans l'univers consumériste* ». C'est pour cela qu'aucun produit n'échappe aux processus d'esthétisation actuellement. L'art a quitté les musées, s'est vaporisé, et, tel un nuage, s'est déposé sur tout notre quotidien sans que l'on y prête garde (Michaud, 2003). L'esthétique tend à teinter notre existence ; cela se voit dans la mode, les cosmétiques, le design, l'architecture, la chirurgie dite esthétique. Ce besoin d'esthétisme s'est progressivement implanté en nous: L'« *Homo aesthéticus* » a succédé à l'« *Homo aeconomicus* ». Apple n'est d'ailleurs pas le dernier à l'avoir compris.

Néanmoins, cette vision du produit a eu longtemps du mal à s'appliquer dans le domaine du numérique à cause des limitations techniques. C'est surtout à partir du développement des interfaces graphiques et de la mise en place d'espaces de rencontre formels[24] que les Designers (dans le sens français[25]) commencèrent à investir le champ des IHMs. Ces derniers mirent

[24] Tels que les workshops SIGCHI biennal « *Designing Interactive Systems* » (DIS), en 1995.

[25] En effet, la plupart des designers français ne se considèrent pas comme de simples «concepteurs » mais comme des artistes qui conçoivent des produits. En anglais, le type de « Design » se fait en fonction du qualificatif qui le

progressivement en avant l'importance de l'esthétisme dans la conception (Masaaki Kurosu & Kashimura, 1995; Noam Tractinsky, 1997). Lors de CHI 2002, un forum fut organisé pour faire rencontrer les membres du SIGCHI[26] et la communauté de l'Institut Américain des Arts Graphiques (AIGA). Puis, en 2003, le SIGGRAPH, le SIGCHI et l'AIGA initièrent la conférence « *Designing for User Experience* » (DUX), qui se positionna sur le champ de la conception visuelle et commerciale, ce qui permit de mettre en avant la vision du Design alors que ce dernier se retrouvait à l'époque souvent mis à l'écart dans les papiers de recherche conventionnelle. De ce fait, la tendance selon laquelle le produit numérique doit dépasser la seule valeur *« utilité* » et « *utilisabilité* » s'amplifia.

La captologie et le marketing de l'expérience

Avec le développement d'Internet à la fin des années 90, le commerce électronique prit son envol. Ainsi, tout naturellement, le marketing s'engouffrera à son tour dans le champ du numérique. Durant cette période, une nouvelle discipline visant à modifier le comportement d'achat se développa : c'est l'informatique persuasive, ou « *Captologie* ». Fogg la définira comme étant « *une tentative non coercitive de changement d'attitudes ou de comportements ou les deux* » (Fogg, 2003). Néanmoins, elle se distingue des techniques de marketing classiques, dans le sens où elle les dépasse, en utilisant des ficelles seulement permises par l'interactivité des dispositifs numériques. Cette approche fut appliquée aux domaines du commerce électronique et des sites d'informations, mais aussi dans le cadre de campagnes de changement d'attitudes et de comportements de la vie de tous les jours, à l'aide de divers dispositifs de communications (Brangier & Bastien, 2010). L'informatique persuasive revisite et prolonge les recherches menées en psychologie de l'influence sociale. Nemery, Brangier & Kopp (2010) proposeront plus tard une liste de CRITERES D'ERGONOMIE PERSUASIVE, distinguant ceux s'appliquant aux aspects statiques (Crédibilité, Privacité, Suggestibilité, Personnalisation, Attractivité) de ceux s'appliquant aux aspects dynamiques de l'interface (Sollicitation, Accompagnement initial, Engagement, Emprise).

En parallèle, on observa dans les pays riches que les consommateurs ne recherchèrent plus à acquérir les objets en tant que tels, mais bien pour les idées, les concepts, les images, les expériences qui y sont liées. On achète des expériences, plus que des objets. D'ailleurs, le quart le plus riche de la population mondiale dépense aujourd'hui presque autant pour acheter des expériences à vivre dans le champ des loisirs que pour acheter des biens et services (Bassani, Sbalchiero, Ben Youssef, & Magne, 2010). L'accumulation matérielle ne séduit plus, malgré des tentatives de marketing produit de plus en plus agressives. Ronald Inglehart (1997) montre ainsi que les sociétés qui ont connu des périodes prolongées de richesses matérielles deviennent de plus en plus intéressées par des valeurs telles que le développement personnel. Il ne s'agit ni plus ni moins qu'une confirmation des intuitions de Maslow (1954) sur le fait que des inspirations post-matérielles peuvent se développer une fois que les aspirations matérielles sont assouvies. Ainsi, si la société des années 80 et 90 était décriée comme consumériste et

précède : « *Web Design, Product Design, Game design, User-Centered Design, User Experience Design, …* » Il y a donc autant de type de « Designer » qu'il y a de façon d'appréhender la conception.

[26] Le SIGCHI ou « *Special Interest Group on Computer-Human Interaction* » est la plus grande association de professionnels, chercheurs ou praticiens, dans le domaine des IHMs.

superficielle, la société d'aujourd'hui pourrait être décrite comme une « *société de l'expérience* » (Schulze, 1993), dont les individus recherchent le bonheur par l'acquisition d'événements de vie positifs. D'ailleurs, des études ont montré que l'achat de produits expérientiels (ex : un concert, un voyage, un restaurant) rendait les gens plus heureux que les achats de produits matériels (ex : un vêtement, un bijou, un équipement stéréo) de même valeur (Carter & Gilovich, 2010; Van Boven & Gilovich, 2003). De même, une étude des indicateurs de l'économie américaine sur le siècle passé montre que les individus accordent une valeur plus élevée pour les expériences que les biens, les commodités ou les services (Pine & Gilmore, 1999).

De ce fait, le domaine du marketing va mettre au point un nouveau moyen de stimuler la consommation, en allant au-delà de la vision matérialiste dominante. L'article pionnier de Holbrook et Hirschmann, en 1982, fit apparaitre la notion de marketing expérientiel. Ces derniers déclarèrent « *qu'après avoir vendu des matières premières, des produits manufacturés et des services, les entreprises doivent apprendre à vendre de l'expérience »,* qu'ils décrivent comme *« la poursuite du fantastique, des sentiments et du plaisir* » (Holbrook & Hirschman, 1982). Cette « *expérientialisation des biens* » (Pine & Gilmore, 1999, p. 14) souligne –comme l'approche décrite dans la partie précédente– l'importance de variables négligées jusqu'alors, telles que les émotions, les sensations, le besoin de plaisir ou d'amusement du consommateur. Elle apparaît aux managers, toujours en quête de sources de différenciation, comme un moyen d'enrober le fonctionnel (Eiglier, 2004) en vue de ré-enchanter la consommation (Firat & Venkatesh, 1995). Depuis lors, le courant expérientiel n'a cessé de prendre de l'importance, au point que l'on n'hésite plus à parler d'un « *véritable paradigme expérientiel* » (Hetzel & Volle, 2002).

Un bon exemple est celui du développement du luxe. En 2012, le chiffre d'affaires du luxe expérientiel (séjours exceptionnels, consommation de denrées rares et coûteuses, …) représentait plus de cinq fois celui du luxe traditionnel (bijoux, ameublement, voiture, …), soit environ mille milliards d'euros (Michaud, 2013). On constate par ailleurs un déplacement des attentes des clients pour les palaces ou séjours haut de gamme. Il ne s'agit plus seulement de proposer un confort luxueux, avec des prestations fortement standardisées, telles que le petit déjeuner américain ou continental, pour ne pas trop bousculer les habitudes culinaires des convives. Il s'agira, au contraire, de proposer au client une atmosphère nouvelle et enrichie, se traduisant par un éveil intellectuel et social, à travers des expériences de vie uniques. Ce

Figure 31 – Vaisseau spatial de touriste Virgin Galatic

Figure 32 – Chambre de l'hôtel de Glace au Canada

nouveau luxe prendra la forme d'un voyage dans l'espace (Figure 31) ou d'une nuit dans un palais de glace (Figure 32). Ces demandes peuvent être totalement excentriques, ou alors essayer de revenir aux sources, avec le courant du « *Rough Luxe* », qui estime que le vrai luxe c'est l'authenticité (Michaud, 2013). Un autre exemple est celui de l'industrie de la musique. Face à la chute des ventes de disques, à cause de la copie sans altérations et sans frais, les industries ont dû se tourner vers des produits non piratables. Et qu'est-ce qui n'est pas reproductible par excellence? Une sensation ou une expérience (Huxley, 2010). C'est pour cela que même si le nombre de ventes d'albums a considérablement chuté, le nombre de concerts ne fait qu'augmenter. Par exemple, la tournée mondiale de Madonna pour « *Confessions on a Dancefloor »* a généré plus de 200 millions de dollars alors que seulement 1,6 million de CD ont été vendus pour cet album.

Dans le domaine de l'IHM, même si la notion d'expérience a été utilisée marginalement avant Norman, c'est ce dernier qui contribua le plus à sa démocratisation. En effet, lorsque Don Norman rejoignit l'entreprise Apple en 1993, en tant que vice-président de la recherche et directeur de l' « *Advanced Technology Group* » (ATG), il apporta avec lui la notion « *d'Expérience Utilisateur* ». Il changea notamment le libellé du poste *« d'Architecte d'Interface Utilisateur* » en *« Architecte d'Expérience Utilisateur* ». Puis, la notion se propagea dans la communauté IHM, suite à un article dans la conférence CHI 1995 « *What You See, Some of What's in the Future, And How We Go About Doing It: HI at Apple Computer* » (Norman, Miller, & Henderson, 1995). À ce sujet, Don Norman déclara : « *J'ai inventé ce terme, car je pensais que "Interface Humaine" et "Utilisabilité" étaient des notions trop étroites. Je voulais couvrir tous les aspects de l'expérience d'un individu avec un système, en incluant le design industriel, le graphisme, l'interface, l'interaction physique et le manuel* ». Le tour de force majeur fut que cette notion fédéra tout un ensemble hétéroclite d'acteurs autrefois séparés (psychologues, agents marketing, designers, informaticiens, …), ce qui permit la naissance d'un nouveau paradigme de recherche en IHM.

Le choc des paradigmes : Utilisabilité vs Expérience Utilisateur

L'agrandissement des possibilités techniques, couplé à divers bouleversements sociétaux, économiques, voire idéologiques, va progressivement fragiliser le paradigme d'évaluation

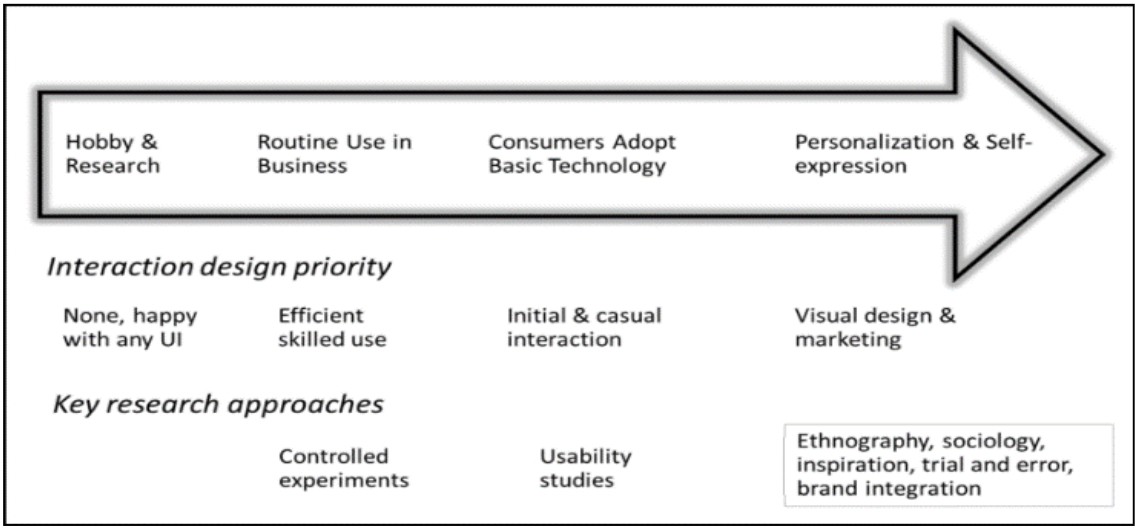


Figure 33 – La technologie de l'invention à la maturité. (*Grudin & Poltrock, 2012*)

dominant, et participer à l'émergence d'un autre, beaucoup plus riche : celui de l'expérience utilisateur.

Au fur et à mesure qu'une technologie nouvelle arrive à maturité, nous devenons de plus en plus exigeants à son égard en y greffant des considérations supplémentaires (Grudin, 2012a). Au commencement, quand la technologie est la seule affaire des chercheurs et des amateurs, ses utilisateurs sont satisfaits par de simples fonctionnalités. Dans ce contexte, l'équipe de développement se concentre sur la résolution des problèmes techniques. Cette phase ne dure jamais très longtemps, car la concurrence voit vite l'opportunité de rentrer sur ce marché émergent. Après un certain temps, et une fois que d'autres compétiteurs rentrent dans l'arène, les fonctionnalités, l'efficience ou la facilité d'utilisation deviennent des facteurs de différenciation. Enfin, quand le marché devient mature, et que la majorité des acteurs ont intégré toutes les fonctionnalités clés de manière satisfaisante, d'autres stratégies de différenciation font leur apparition, comme la personnalisation ou l'esthétisme (Figure 33). De plus, cette forte pression du marché pousse les entreprises à proposer leurs produits sur un mode de développement très réactif. L'extension du nombre de facteurs à prendre en compte lors de la conception et la réactivité demandée du marché poussent les entreprises à opter pour des méthodes d'évaluation rapides et agiles, plus encore que lors de la révolution des METHODES D'UTILISABILITE DISCOUNT de Nielsen.

Cette tension grandissante conduira, en 1998, à un choc de paradigmes d'évaluation, à la suite d'une série d'études provocatrices. En effet, en 95, durant la conférence de CHI, Wayne Gray exposa une critique violente des études certifiant la validité des METHODES D'EVALUATION DISCOUNT, lors d'une présentation appelée : « *Discount or Disservice ? Discount usability analysis - evaluation at a bargain price or simply damaged merchandise ?* ». Le niveau d'intérêt dans la discussion amena Wayne Gray et sa collaboratrice Marilyn Salzman à conduire une enquête plus systématique sur cinq grandes études publiées sur la validation d'une variété d'UEMs (« *USABILITY EVALUATION METHODS* »). L'article de recherche résultant fut publié en 1998. Il dénonça et démontra les failles des études censées justifier l'utilisation de nombreuses méthodes d'évaluation, telles que l'EVALUATION HEURISTIQUE, le COGNITIVE WALKTHROUGH, les TESTS UTILISATEURS (Gray & Salzman, 1998). L'article reprendra donc cinq études faisant référence dans le domaine (Desurvire, Kondziela, & Atwood, 1993; Jeffries, Miller, Wharton, & Uyeda, 1991; Karat, Campbell, & Fiegel, 1992; Nielsen & Phillips, 1993; Nielsen, 1992) et étudiera leurs validités selon la taxonomie de Cook & Cambell (1979b), c'est à dire selon leur validité statistique, interne, externe et de construit. Les auteurs arrivèrent à un verdict sans appel : les études sur les méthodes d'utilisabilité souffrirent de problèmes sur les quatre types de validité, ce qui invalida leurs conclusions et ne permit pas de connaître la véritable fiabilité des méthodes d'évaluation d'utilisabilité examinées. En conclusion, ces derniers suggérèrent de corriger ces biais en procédant à un approfondissement des protocoles d'évaluation initiaux.

Cet article relança la controverse entre deux camps existant depuis longtemps : les chercheurs « *rigoureux* » et les praticiens « *pragmatiques* ». Néanmoins, les germes d'une position nouvelle avaient également émergé, poussés par la plus grande diversité académique à la fin des années 90. La relecture de l'article impliqua de nombreux chercheurs et les commentaires associés furent si importants qu'un numéro spécial du célèbre Journal « *Human-Computer Interaction* » lui fut dédié afin de tous les compiler et d'apporter un « *droit de réponses* » aux attaques en

règles. Les réactions à cet article furent compilées dans le papier fleuve de 61 pages « *Commentary on "Damaged Merchandise ?"* », édité par Olson & Moran (1998). Dix commentaires, effectués par différents acteurs (praticiens en contexte industriel, managers de conception, experts en méthodologie, chercheurs en design) furent ainsi exprimés en guise de réponse. Il est important de reprendre chacun de ces commentaires pour comprendre la diversité des raisons qui poussèrent ces nouveaux acteurs à s'opposer à la vision « *Expérimentaliste / Cognitive* », toujours dominante dans la communauté de recherche en IHM de l'époque :

- **Commentaire 1 – « *The Fine Art of Comparing Apples and Oranges* » (John Kant) :** John Kant défendit l'utilité des UEMs dans le contexte historique de l'époque. En effet, dans les années 90, l'intérêt pour utiliser ces nouvelles méthodes était grand et il valait mieux dire « *quelque-chose* » à leur sujet, tôt, que « *quelque-chose plus sûr*» à leur sujet, tard.
- **Commentaire 2 – « *Ivory Towers in the Trenches: Different Perspectives on Usability Evaluations* » (Robin Jeffries and James R. Milk)** : Ces deux chercheurs défendirent l'utilisation de données écologiques plutôt que des protocoles expérimentaux idéalisés. Pour eux, pour comprendre un sujet si complexe, il valait mieux avoir une vue d'ensemble de la question, avec un enracinement solide au réel, même aux dépens de la pureté expérimentale. Néanmoins, Jeffries et Milk ne privilégièrent aucune des deux approches et précisèrent que toute entreprise de recherche est un compromis entre la rigueur expérimentale et le bénéfice d'étudier un phénomène dans son contexte.
- **Commentaire 3 – « *Damaged Merchandise? Comments on Shopping at Outlet Malls*» (Arnold M. Lund)** : Lund attira le regard sur l'importance de la réalisation de ces études dans un contexte industriel car cela les rapprochaient plus des pratiques réelles. En effet, pour Lund, la valeur d'une recherche réside dans sa validité apparente (« *face validity* ») plutôt que par sa pureté méthodologique. Le monde de l'industrie n'a pas besoin de réponse parfaite, mais plutôt de recommandations pragmatiques permettant l'amélioration des pratiques existantes.
- **Commentaire 4 – « *Damaged Merchandise: How Might We Fix It?* » (Ian McClehnd) :** McClelland mit l'accent également sur la différence de priorité entre les praticiens et les chercheurs dans le domaine des IHMs. Néanmoins, il ajouta un commentaire intéressant sur la nécessité de dépasser la conception classique d'utilisabilité. En effet, pour l'auteur, l'utilisabilité doit être vue comme une créature multidimensionnelle qui doit changer de nature en fonction des circonstances. Dans ce contexte, le critère décisif de toute étude d'utilisabilité est d'identifier les facteurs critiques pour le succès et l'utilisation satisfaisante d'un produit. Il parut clair pour l'auteur que les évolutions récentes, que ce soit dans le domaine des produits professionnels ou dans celui de la consommation, diversifièrent les facteurs contribuant au succès d'un produit (ex : les émotions et le plaisir)
- **Commentaire 5 – « *A Case for Cases* » (Bonnie E. John)** : John pensa que la position de Gray & Salzman sur la validation expérimentale était excessive et recommanda à la place l'utilisation d'étude de cas car les UEMs étaient encore dans une étape formative de développement. En effet, l'auteur déclara qu'il vaut mieux utiliser des méthodes d'évaluation permettant un recueil d'information riche, car l'immaturité des UEMs rend encore nécessaire une investigation qualitative et globale du sujet.

- **Commentaire 6 – « *Experiments Are For Small Questions, Not Large Ones Like "What Usability Evaluation Method Should I Use? (*Andrew E. Monk*)*** : Monk commença par prendre du recul et décrire le rôle de l'expérimentation dans le champ de la psychologie, là où les questions posées sont très précises. Puis, il décrira pourquoi cette pratique ne peut pas donner de réponse sur des interrogations plus larges et de surcroît pour des problèmes aussi complexes que l'évaluation des UEMs. En effet, pour lui, le rythme d'innovation (rendant caduques les réplications) et le nombre de facteurs à prendre en compte rendaient la tâche de l'expérimentateur impossible pour ce genre de questions.
- **Commentaire 7 – « *What's Science Got To Do With It? Designing HCI Studies That Ask Big Questions and Get Results That Matter*» (Sharon L. Oviatt)**: Oviatt, lui, n'etait pas contre une entreprise expérimentale ayant pour objet l'évaluation des UEMs. Néanmoins, celle-ci ne devait pas se contenter de comparer les UEMs entre elles, car ayant chacune quelque chose de différent à dire sur l'utilisabilité du produit. Il posa une question plus intéressante pour lui : « *Comment combiner ces méthodes efficacement pour améliorer l'évaluation d'une application ?* » De plus, l'auteur déclara que pour contribuer à la science, les chercheurs en IHMs devaient entreprendre un programme systématique de grande ampleur, longitudinal, avec des études de réplication, pour régler les problèmes complexes qui se posent lors de l'évaluation des UEMs.
- **Commentaire 8 – « *Review Validity, Causal Analysis, and Rare Evaluation Events* » (John M. Carroll)** : Carroll critiqua également l'utilité des méthodes expérimentales dans le champ de l'étude de l'utilisabilité car considérées trop « étroites » pour les IHMs. Pour cet auteur, il y a un compromis à faire entre la rigueur et la pertinence d'une étude. Il ajouta même que ce sont souvent les études les plus pertinentes qui sont les plus influentes et non pas les plus rigoureuses. Il conclut en disant que pour évaluer les UEMs, il convient d'utiliser des approches complémentaires.
- **Commentaire 9 – « *Triangulation Within and Across HCI Disciplines* » (Wendy E. Mackay) :** Mackay mit en avant l'utilisation de méthodes d'évaluation multiples, triangulées à l'intérieur et à l'extérieur des différentes disciplines IHMs pour déboucher sur une vue plus globale de la question. Pour l'auteur, aucune étude isolée ne pourra prévenir toutes les formes de biais de validité en même temps. De ce fait, la diversité des méthodes employées pour une même entreprise de recherche est un gage solide de confiance pour la validité des résultats obtenus.
- **Commentaire 10 – « *On Simulation, Measurement, and Piecewise Usability Evaluation* » (William M. Newman) :** Newman souligna le rôle primordial de la qualité de la simulation des situations futures lors de l'évaluation. En effet, un test utilisateur, même coûteux, peut être invalidé si la situation d'intérêt est simulée trop étroitement. Cet auteur conseilla de prendre en compte le contexte du système le plus large possible et de ne pas fragmenter l'activité de l'utilisateur. Les études combinant une approche analytique (centrée sur les tâches) et une approche ethnographique, permettraient ainsi d'accéder à la structure globale de l'activité.

Ainsi, nous voyons, grâce à l'ensemble de ces réactions, que le différend sur l'évaluation des méthodes dépassa la simple question méthodologique : il devint épistémologique. En effet, un système de recherche est considéré valide si le mode de production de la connaissance qui le

sous-tend est considéré comme valide ; et nous avons vu, par ailleurs, qu'il y a eu de grandes différences sur la manière de trancher cette question au fil de l'histoire des IHMs. La façon dont les évaluations doivent être menées dépend grandement des perspectives disciplinaires qui composent les IHMs à un moment donné. Quand de nouveaux aspects de l'interaction prennent de l'importance, de nouvelles disciplines sont intégrées pour traiter ces sujets inédits. Ces dernières n'apportent pas seulement avec elles de nouvelles méthodes, mais également d'autres idées sur la nature de la connaissance et les manières de la construire (Kaye & Sengers, 2007).

Nous voyons ainsi réapparaître, dans une mesure plus profonde encore, le conflit des paradigmes d'évaluation entre expérimentateurs réductionnistes et praticiens holistes, tel qu'il était déjà apparu lors de la scission des IHMs et des facteurs humains dans les années 80. Il s'agira, en somme, d'une bataille entre les défenseurs de la précision, permise par le contrôle expérimental, et de la validité écologique, permise par l'étude en contexte. Néanmoins, à l'approche des années 2000, le courant de plus en plus qualitatif prôné par les praticiens, et l'élargissement du domaine à de nouvelles disciplines, telles que le Design ou le marketing, affaiblit d'autant plus les rapports de force des positions rigoristes et cognitivistes, encore très influentes dans les processus de publication des articles et des actes de conférence dans le milieu des IHMs. À partir de ces commentaires, dont l'influent journal « *Human-Computer Interaction* » permit une large diffusion, trois constats fondamentaux se propagèrent. Le premier stipula, qu'en conséquence de l'évolution du domaine, il fallait aller au-delà de l'utilisabilité, en enrichissant l'évaluation d'autres facteurs devenus tout aussi importants. Deuxièmement, la méthodologie utilisée ne devait plus se limiter aux procédures expérimentales car toute méthode a ses limites. Troisièmement, c'est en triangulant ces différentes approches qu'il est possible de contrebalancer leurs limites respectives. Une nouvelle crise paradigmatique fut ainsi mise sur le devant de la scène et agita alors toutes les communautés. Green et Jordan (1999) déclarèrent un an plus tard : « *L'ergonomie/facteurs humains se trouve à un moment crucial de son développement. C'est une discipline qui a toujours été liée au processus de conception du travail, des systèmes techniques et des produits... Cependant, la révolution électronique a mis au jour une quantité impressionnante de problèmes qui étaient considérés jusqu'ici comme "ésotériques" dans le domaine de l'ergonomie... apportant de nouvelles demandes aux ergonomes et à ceux qui interviennent dans les processus de conception* » (p. 249). Un nouveau paradigme va alors émerger, celui de « *l'Expérience Utilisateur* » (Harrison et al., 2007), dont la portée va progressivement fédérer tous les différents acteurs du domaine.

La formation et consolidation du domaine de l'UX

Le paradigme de l'Expérience Utilisateur (UX) est « *né de l'influence réciproque de l'évolution des technologies, de l'introduction de ces dernières dans tous les secteurs de l'activité humaine (travail, loisir, etc.) et de la mise à disposition de ces technologies à des utilisateurs de plus en plus variés et ayant des besoins de plus en plus diversifiés. La satisfaction des besoins des utilisateurs en termes de fonctionnalités a permis de faire apparaître les besoins en termes de facilité d'utilisation, puis de divertissement, puis de contacts sociaux* » (Brangier & Bastien, 2010, p. 20). Depuis ces quinze dernières années, ce nouveau paradigme de recherche s'est développé en passant par toute une série de phases de maturation. Afin de cerner précisément

les bouleversements actuels dans le champ des IHM, il convient de parcourir ces différentes étapes d'élaboration. La première a été de se mettre d'accord sur une définition du concept, ses caractéristiques et en quoi ce dernier se différencie fondamentalement de l'utilisabilité. Puis, à partir de ces notions, différentes approches méthodologiques vont émerger pour diriger les recherches dans ce nouveau paradigme en IHM.

Le rapprochement des communautés en vue d'une unification du concept d'expérience utilisateur

Pour Hassenzahl et Tractinsky (2006), les tendances de recherche dans le domaine de l'expérience utilisateur ont d'abord été prospectives dans les années 90, conceptuelles début 2000, pour enfin devenir empiriques au milieu des années 2000. L'utilisation du terme se diffusa à partir de sa première utilisation « publique » en 1995, pour exploser durant les années 2000 (Figure 34).

L'expérience utilisateur est apparue au début comme une notion « *parapluie* » dont la fonction a été l'émergence de nouveaux moyens de compréhension et d'étude de la qualité de l'interaction numérique (Bargas-Avila & Hornbæk, 2011). Les premières recherches UX au début du millénaire ont consisté à étudier, prospectivement, et de manière isolée, des variables mises de côté jusqu'alors dans l'interaction, tels que le plaisir (Jordan, 1998; Jordan, 2000), la joie (Hassenzahl, Beu, & Burmester, 2001) ou les qualités hédoniques d'un produit (Hassenzahl, Platz, Burmester, & Lehner, 2000). Les recherches suivantes ont consisté à connecter toutes ces approches divergentes en recherchant à définir, conceptualiser, et caractériser leur dénominateur commun : « *l'Expérience Utilisateur* ». Il s'agira d'une étape fondamentale pour plusieurs raisons (Roto, 2007) : (i) une définition commune facilite le débat scientifique, en particulier lorsque des chercheurs de plusieurs disciplines sont impliqués; autrement, des problèmes de communication se produisent inévitablement; (ii) elle permet la gestion des applications pratiques de UX, son opérationnalisation et sa mesure. (iii) Elle aide à l'enseignement de la notion d'UX par la compréhension de ses fondements et de sa portée.

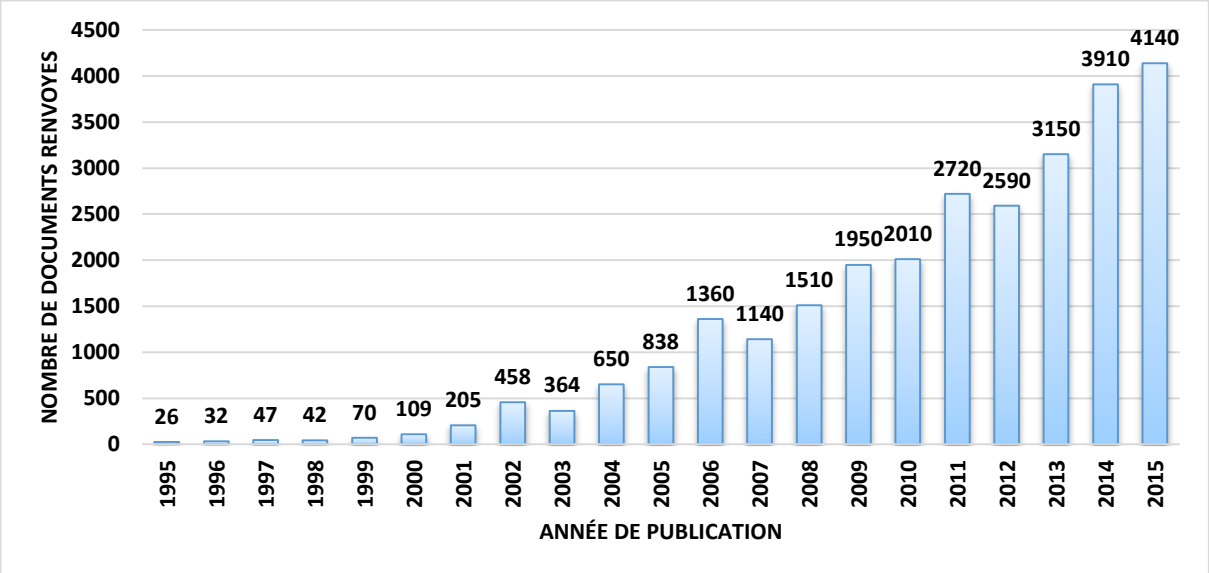


Figure 34 – Nombre de documents par année renvoyés par scholar.google.com avec la requête « "user experience"+HCI » entre 1995 et 2014 (effectué le 07/04/2016)

Néanmoins, la mise au point d'une définition universelle de l'UX est un exercice particulièrement difficile pour plusieurs raisons (Lai-Chong Law, Roto, Hassenzahl, Vermeeren, & Kort, 2009). Premièrement, l'UX est associé à un large panel de concepts, dynamiques et flous, incluant des facteurs émotionnels, expérientiels, hédoniques ou encore esthétiques. L'inclusion ou l'exclusion d'une de ces variables est un choix arbitraire, qui dépend du parcours du chercheur et de ses intérêts. Deuxièmement, l'unité d'analyse de l'UX est très variable, allant d'un seul aspect de l'interaction-utilisateur, avec une seule application à une vision globale, multi-utilisateurs, multi-applicative, et multidisciplinaire (Sward, 2006). Troisièmement, le paysage de la recherche UX est très fragmenté et rendu complexe par la diversité des modèles théoriques dont, chacun, aborde le sujet différemment.

Il a donc fallu organiser en premier des espaces de rencontre et d'expression multidisciplinaire afin de consacrer du temps et un espace spécialement sur le sujet. Cela se fit sous la forme de workshops, meetings, forums, conférences et numéros spéciaux de revues. Par exemple, en 2004, Donald A. Norman, déplorant le manque de dialogue entre les disciplines, proposa une section spéciale dans la revue « *Human-Computer Interaction* », dans un numéro de décembre 2004, nommé : « *Beauty, Goodness, and Usability* ». Norman était intéressé par la publication du débat autour de l'article de Noam Tractinsky et collaborrateurs, au titre provocateur de « *What is Beautiful is Usable* » (Tractinsky, Katz, & Ikar, 2000), dans l'objectif de faire avancer la discussion sur les liens théoriques entre utilisabilité et esthétisme (Hassenzahl, 2004b; Norman, 2004b). De plus, fin 2004, un meeting BayCHI sur le sujet « *Expérience Utilisateur : Pourquoi tant d'organisations pensent qu'elle leurs appartient ?* » fut organisé entre Norman et un ensemble de représentants de ces organisations[27]. Ces derniers arrivèrent à la conclusion évidente que la propriété de l'UX est partagée entre tous. D'autres espaces de rencontres multidisciplinaires et d'échanges sur l'UX furent organisés, tel qu'un consortium de développement (DevCon), appelé « *Meeting the Needs of the Multidisciplinary Professional and of the Multiple Professional Associations and Events of Importance to Them* » lors de CHI 2005 ; ou encore un numéro spécial du magazine *Interaction* « *Whose profession is it anyway ?* », édité par Pabini Gabriel-Petit (2005), le webmaster du site UXmatters. Dans ce contexte, il déclara que « *dans une culture de collaboration plutôt que de compétition, la pollinisation croisée des idées entre toutes les disciplines professionnelles est une attitude saine* » (p.17). Le rapprochement des communautés se consolida et leurs travaux autour de l'UX se structurèrent. De nombreuses contributions visant à unifier le champ de l'UX apparurent dans la série de conférences DUX, CHI, NordCHI et même à l'intérieur du projet de coopération européen COST294-MAUSE[28] (Law, Hvannberg, & Hassenzahl, 2006; Law, Vermeeren, Hassenzahl, & Blythe, 2007; Law, Roto, Vermeeren, Kort, & Hassenzahl, 2008; Law et al., 2009).

---

[27] Les organisations représentées etaient, l'*AIfIA* (*Asilomar Institute for Information Architecture*), l'*AIGA Experience Design*, le *BACHFES* (*Bay Area Chapter of the Human Factors and Ergonomics Society*), le *BayCHI*, la branche *IDSA* (*Industrial Designers Society of America*) de San Francisco, l'*IxDG* (*Interaction Design Group*), la branche *SIGGRAPH* (*ACM Special Interest Group on Computer Graphics and Interactive Techniques*) de San Francisco et de la Silicon Valley, la *STC* (*Society for Technical Communication*), la *UPA* (*Usability Professionals Association*) et l'*UXnet* (*User Experience Network*).

[28] Dont les delivrable sont téléchargeables ici : http://www.cost294.org/

Premier effort de caractérisation de l'expérience utilisateur

La première façon d'initier le travail, la plus intuitive, a été de cerner en quoi l'UX se différencie de l'utilisabilité. Hassenzahl (2006) va se prêter à l'exercice en proposant trois distinctions fondamentales de nature. Il discernera l'UX par ses caractéristiques (a) holistique, (b) subjective et (c) positive.

(a) Hassenzahl distingua en premier l'UX de l'utilisabilité par sa nature **holistique**. En effet, l'utilisabilité se concentre essentiellement sur les tâches de l'utilisateur et leurs accomplissements, ce qui correspond à une focalisation sur la sphère **pragmatique** de l'interaction ; au contraire, l'expérience utilisateur prône une approche plus **holistique**, c'est-à-dire en recherchant un équilibre entre la satisfaction de buts **pragmatiques** mais également de buts non orientés tâches (cf. **hédoniques**), tels que l'esthétisme, la stimulation ou encore le divertissement

(b) Deuxièmement, l'expérience utilisateur sera vue, par Hassenzahl, comme **subjective**. En effet, le domaine des IHM ayant hérité d'une approche issue en grande partie de la psychologie expérimentale, elle a tendance à estimer principalement les techniques de recueil permettant de générer des données **objectives** et à éviter tant que possible les autres. Dans le champ de l'expérience utilisateur, l'attrait pour les données subjectives est renforcé, voire favorisé. En effet, c'est l'appréciation subjective des utilisateurs qui va guider leur comportement, et donc qui va définir la valeur d'un produit. En d'autres mots, c'est la valeur perçue d'un produit, la façon dont l'expérience subjective va être éprouvée qui est importante, et non pas sa valeur « *objective* » en soi ; bien qu'il soit intéressant de connaitre les mécanismes permettant aux qualités objectives d'un produit de se traduire en appréciation subjective positive.

(c) Enfin, l'expérience utilisateur se distinguera, pour Hassenzahl, par sa nature **positive**. L'utilisabilité, en se focalisant sur la suppression de problèmes, du stress ou de la frustration, cherche à diminuer les conséquences **négatives** d'un produit. A l'inverse, l'expérience utilisateur tente de maximiser les aspects **positifs** de celui-ci, en essayant par exemple de générer de la joie, de la fierté ou en favorisant certains aspects fortement appréciés. Entre autres, on constate que les hauts degrés de satisfaction ne peuvent s'atteindre que si le produit possède des qualités motivationnelles intrinsèques, permettant par exemple la reconnaissance de soi, l'accomplissement ou le challenge. Même la meilleure utilisabilité au monde aura du mal à dessiner un sourire sur le visage d'un utilisateur (Timo, 2005). Ainsi, de nombreux chercheurs souhaitèrent, comme Seligman et Csikszentmihalyi (2000) le firent pour la psychologie positive, porter le courant UX vers IHM positif, c'est à dire vers la création d'expériences de qualité plutôt que vers la prévention de problèmes d'utilisabilité (Hassenzahl & Tractinsky, 2006).

La communauté des IHM, convaincue dans une large majorité de l'utilité et de la nécessité de résoudre les défis nouveaux soulevés par l'expérience utilisateur, commença à développer un agenda de recherches UX en trois grands thèmes (Figure 35). Un premier se focalisa sur les besoins **non-instrumentaux** des utilisateurs (i) ; un second traita des aspects **émotionnels** et **affectifs** de l'interaction (ii) ; et le dernier se préoccupa plus particulièrement de la nature **expérientielle** de l'UX (iii) (Hassenzahl & Tractinsky, 2006).

(i) Nous avons vu que, depuis ses débuts, les recherches en IHM se sont concentrées sur les méthodes permettant d'augmenter l'efficacité et l'efficience des utilisateurs dans un environnement de travail donné. De ce fait, la valorisation des **qualités instrumentales** d'un produit a constitué depuis longtemps l'effort principal du domaine. Cependant, cette vision étroite fut de nombreuses fois débattues. L'esthétisme (Alben, 1996), la confiance (Cody-Allen & Kishore, 2006) ou l'intimité (Gaver & Martin, 2000) ont été des exemples de besoin non-instrumentaux ayant un impact direct sur la satisfaction de l'utilisateur. Bevan (2008) pensa ainsi que la satisfaction des utilisateurs dépend de la validation combinée des besoins **pragmatiques** et **hédoniques**. Les **besoins pragmatiques** de l'utilisateur se caractérisent par une expérience d'utilisation perçue acceptable (incluant l'efficience), des résultats perçus acceptables (incluant l'efficacité) et des conséquences d'utilisation perçues acceptables (incluant la sécurité). Hassenzahl (2003) identifia trois **besoins hédoniques**, qui sont la stimulation (développement personnel en termes de compétences ou de connaissances), l'identification (expression de soi, interaction avec les autres) et l'évocation (identité personnelle, souvenirs). Bevan (2008) ajouta un quatrième besoin hédonique qui est la réaction émotionnelle positive vis-à-vis du produit (le plaisir viscéral de Norman , 2004). Ces différentes approches permirent d'enrichir les modèles actuels de qualité des produits pour permettre une évaluation plus holistique (Hassenzahl & Tractinsky, 2006). Cela conduisit certains auteurs à étudier l'expérience utilisateur sous le prisme de modèles multidimensionnels. Pour Robert et Lesage (2011), l'expérience utilisateur résulte de la combinaison de différents types d'expériences perçus par l'utilisateur envers le produit (fonctionnel, physique, perceptuel, cognitif, psychologique et social), avec deux méta-niveaux, le sense-making (fabriquer ou donner du sens) et l'esthétique. De plus, il existe des interactions entre ces différentes dimensions, qui peuvent se révéler complexes. Par exemple, Diefenbach et Hassenzahl (2009) ont montré que la beauté des produits est une dimension usuellement valorisée, mais qui est mise de côté quand il s'agit de choisir entre un téléphone portable seulement beau ou utilisable. C'est pour cela que les efforts actuels de recherche visent à améliorer la modélisation de l'expérience utilisateur (Law & Van Schaik, 2010; Van Schaik & Ling, 2008), afin d'améliorer son évaluation (modèles de mesure) et de cerner la nature des liens entre ses dimensions (modèle structuraux) ; (Edwards & Bagozzi, 2000). D'un point de vue opposé, d'autres auteurs ont conclu que le caractère holistique de l'UX impliquerait au contraire une insécabilité de celle-ci, ce qui l'oblige donc à être pensé et étudié comme tel (McCarthy & Wright, 2004).

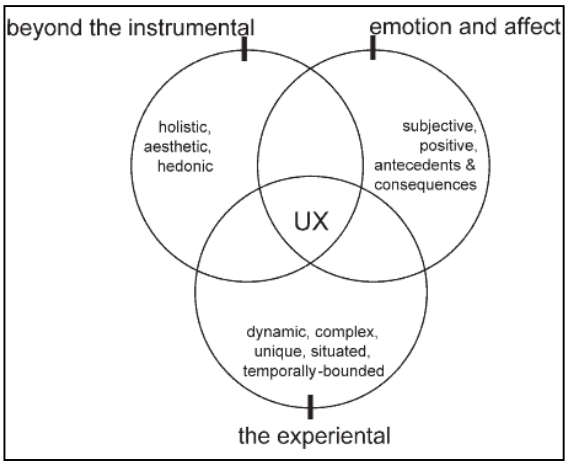


Figure 35 − Les trois axes de recherche de l'expérience utilisateur (Hassenzahl et al., 2006)

(ii) Les recherches actuelles soulignent l'impact du **système émotionnel** sur une large gamme de processus mentaux tels que la prise de décision (Loewenstein & Lerner, 2003) ou le sentiment de bien-être (Suh, Diener, & Fujita, 1996). Nous avons déjà vu que Picard fut un des premiers chercheurs à s'intéresser aux émotions dans le cadre des IHM (Picard, 1997). Il

s'agissait de détecter les émotions des utilisateurs −majoritairement négatives− pour gérer leurs frustrations ou prévenir d'autres émotions déplaisantes. Cependant, même si les chercheurs en UX partagèrent une méthodologie commune pour la capture des émotions, ils intéressèrent par les conséquences des affects, du côté humain, que de l'extension de l'appareillage, du côté technologique (Hassenzahl et al., 2006). En ce sens, ils étaient plus concernés par la compréhension du rôle des affects comme antécédents, conséquences et médiateurs lors de l'utilisation d'une technologie. Ils se concentrèrent également davantage sur les **émotions positives** telles que la joie, le fun ou la fierté (Hassenzahl et al., 2006). C'est dans cette voie que les émotions constituent une branche de recherche particulièrement intéressante pour l'évaluation de l'expérience utilisateur. En effet, malgré la difficulté conceptuelle pour les appréhender, ces dernières permettent d'évaluer intuitivement le plaisir et la douleur, en fournissant « *l'étalon sur lequel des possibilités qualitativement différentes peuvent être comparées*» (Russell 2003, p. 153). Les études sur les états subjectifs traduisant une bonne expérience utilisateur seront d'un grand intérêt pour l'évaluation de l'UX. Elle se poursuivent aujourd'hui au travers des notions telles que le « *Flow* » (Chen, Wigand, & Nilan, 2000; Pace, 2004; Procci, Singer, Levy, & Bowers, 2012; Rodríguez-Sánchez, Schaufeli, Salanova, & Cifre, 2008; Sherry, 2004), l'immersion (Fornerino, Helme-Guizon, & Gotteland, 2006; Jennett et al., 2008) ou encore l'engagement (Boyle, Connolly, Hainey, & Boyle, 2012; Higgins & Scholer, 2009; O'Brien & Toms, 2008).

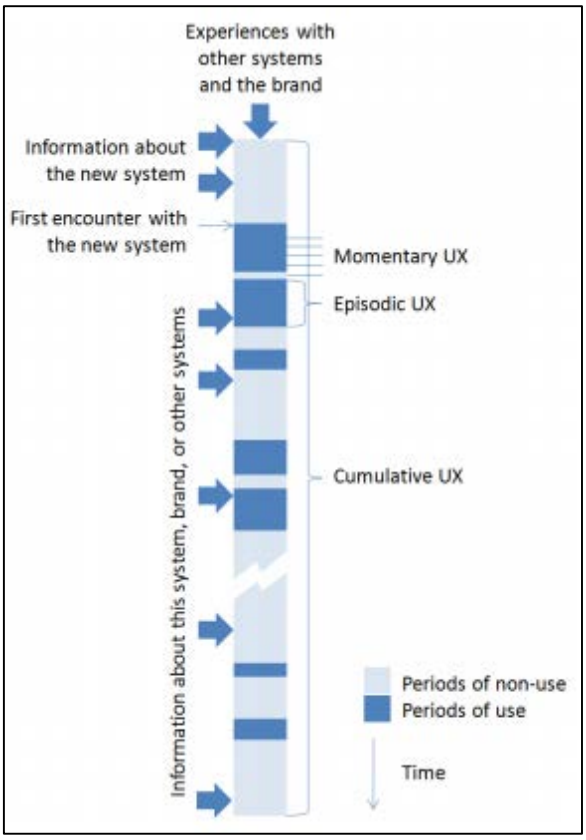


Figure 36 – L'UX en fonction de la temporalité ***(Roto, Law, Vermeeren, & Hoonholt, 2011)***

(iii) Enfin, les recherches sur l'UX commencèrent à explorer sa dimension « **expérientielle** ». Cette perspective souligne deux aspects en particulier de l'utilisation d'une technologie : sa localité et sa temporalité. De ce point de vue, une expérience est une combinaison unique de plusieurs éléments : l'environnement immédiat, le système numérique et les états internes de l'utilisateur (humeurs, attentes, buts), qui évoluent conjointement au cours du temps, en possédant un début et une fin (Robert & Lesage, 2011). De ce fait, c'est l'utilisation d'un produit dans un certain contexte et à un certain moment qui va générer une expérience particulière. Cela pose également la question de la temporalité appropriée pour étudier l'UX. D'une extrémité, nous pouvons nous concentrer sur l'expérience d'un très court moment, telle qu'une réponse viscérale durant l'usage ; de l'autre, nous pouvons prendre en compte toute une série d'expériences cumulatives, formées par une alternance d'épisodes d'usage et de non-usage, sur une période de plusieurs mois ou plus encore. L'expérience peut être ainsi saisie comme un changement bref d'affect durant l'interaction (**UX momentanée**), l'évaluation d'un épisode

d'usage spécifique (**UX épisodique**) ou encore durant une certaine période de temps (**UX cumulative**) ; (Figure 36). Rien n'empêche également d'étudier l'UX avant l'usage (**UX attendu**) ou en sa totalité (**UX globale**) ; (Roto, 2007). Dans cet ordre d'idées, Karapanos, Hassenzahl, & Martens (2008) évaluèrent comment les utilisateurs forment des jugements évaluatifs au cours du temps vis-à-vis d'un produit. Ils constatèrent que lors de la première expérience d'interaction, l'évaluation de la qualité d'un produit est formée en grande partie par les aspects pragmatiques de celui-ci (utilité et utilisabilité) ; alors qu'après quatre semaines, ce sont plutôt certaines qualités hédoniques, telle que l'identification, qui deviennent un des aspects dominants de sa qualité perçue. De plus, ils constatèrent que l'esthétisme, qualité très attrayante lors de la première expérience avec le produit, perdit progressivement de sa force au cours du temps. De la même manière, Karapanos, Zimmerman, Forlizzi et Martens (2009) suivirent six personnes ayant acheté un iPhone™ et constatèrent que leurs motivations passèrent progressivement de besoins plutôt hédoniques à des aspects plus profonds, plus identitaires et porteur de sens (identification). Enfin, d'autres chercheurs pensent que l'évaluation rétrospective ne représente pas l'épisode expérientiel total, mais seulement son souvenir le plus récent ; de ce fait, ils proposent l'usage de stratégie de mesure supplémentaire, comme des mesures répétées durant l'épisode expérientiel (Hassenzahl & Sandweg, 2004). Ces recherches soulignent donc le caractère subjectif, temporel et dynamique de nos expériences d'interaction avec les technologies.

Pour Hassenzahl et al. (2006), l'intérêt fort des recherches récentes pour l'expérience utilisateur n'est pas une coïncidence. Il résulte d'un milieu technologique assez mature pour aller au-delà des simples fonctionnalités, d'une demande accrue des utilisateurs en dehors du cadre strictement professionnel et de l'intérêt renouvelé de la communauté scientifique à propos du système affectif et de ses relations avec la cognition. Ces avancées conceptuelles poussèrent les chercheurs vers un effort de définition plus stricte de l'expérience

Convergence des définitions et standardisation de l'expérience utilisateur

Au fil du temps, compte tenu des avancées théoriques du domaine, des chercheurs se sont donnés comme objectif de définir plus formellement l'expérience utilisateur. Cela était d'autant plus important que de nombreux auteurs avaient déjà donné leurs définitions de l'expérience utilisateur, associant l'UX à des significations fort différentes (Forlizzi & Battarbee, 2004). De plus, ces dernières varièrent en matière de longueur, portée et granularité, de la simple mention « *d'action motivée* » à une description détaillée avec des exemples d'états psychologiques utilisateurs et des caractéristiques techniques des systèmes.

Une des premières définitions a été donnée par Alben (1996) par : « *tous les aspects de l'utilisation d'un produit : la sensation dans les mains, à quel point on comprend comment cela marche, ce que l'on ressent quand on l'utilise, à quel point il sert nos objectifs, et à quel point celui-ci est adapté au contexte d'utilisation* ». Pour Shedroff (2001), l'expérience utilisateur est « *la sensation de l'interaction avec un produit, un système, un service, ou un événement, au travers de tous les sens, au cours du temps, et à un niveau à la fois physique et cognitif. Les bornes d'une expérience peuvent être expansives et inclurent le sensoriel, le symbolique, le temporel, et la création de sens* ». Pour Mäkelä & Fulton Suri (2001), l'UX est « *le résultat d'une action motivée dans un certain contexte. Les expériences utilisateur précédentes et les*

attentes modifient l'expérience présente ; l'expérience en cours conduit à plus d'expériences et à une modification des attentes ». Pour Preece, Rogers, & Sharp (2002), cela désigne simplement ce que l'utilisateur « *éprouve* » lors d'une interaction. Enfin, pour Hassenzahl & Tractinsky (2006), l'expérience utilisateur est « *une conséquence de l'état interne de l'utilisateur (prédisposition, attente, besoin, motivation, humeur, etc.), des caractéristiques du système conçu (complexité, objectif, utilisabilité, fonctionnalité, etc.) et du contexte (ou environnement) à l'intérieur duquel l'interaction se passe (milieu organisationnel/social, activité plus ou moins porteuse de sens, usage plus ou moins volontaire, etc.)* ». Cette liste est loin d'être exhaustive. En effet si l'on ne s'en tient rien qu'aux définitions fournies par Roto et al. sur leur site internet[29] nous pouvons en dénombrer au moins 27.

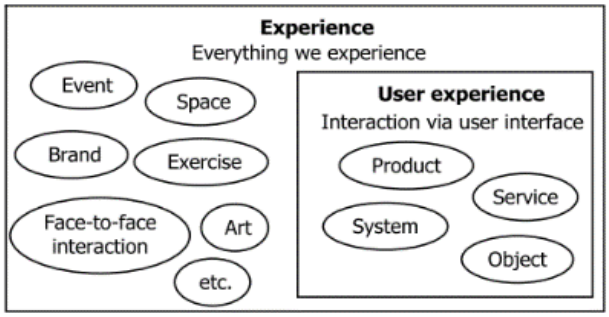


Figure 37 – L'UX en relation avec les autres expériences *(Law et al., 2009)*

Pour permettre la mise au point d'une définition commune, Law et al. (2009) interrogèrent 275 chercheurs et praticiens, issus de domaines académiques et industriels, sur leur vision de l'UX. Malgré quelques divergences, ces derniers arrivèrent à dégager un consensus sur sa nature et sa portée : l'expérience utilisateur est considérée pour la majorité comme **dynamique**, **dépendante du contexte** et **subjective**. Sur les points plus controversés, les auteurs essayèrent de trancher en fonction des données obtenues. Par exemple, sur la plus grande controverse (i) « *l'expérience utilisateur est-il quelque chose d'individuel ou social* ? », les auteurs la définirent comme individuelle, car, même si d'autres personnes peuvent influencer une expérience, seul un individu peut l'avoir et la ressentir (Law et al., 2009). Ces derniers ne nièrent pas l'existence de travaux sur l'expérience utilisateur « *sociale* », mais le situèrent en dehors du champ de l'expérience utilisateur classique. Dénommée « *co-expérience* », les chercheurs dans ce domaine cherchent à comprendre comment cette expérience est co-construite et partagée dans un contexte d'interactions sociales. Pour eux, les enjeux qui portent sur la définition, la théorisation, la qualification et la quantification de la co-expérience, ne se résument pas seulement à la somme des expériences utilisateurs individuels (Battarbee, 2003; Forlizzi & Battarbee, 2004). (ii) Un autre sujet de discorde est la relation qu'entretient l'expérience utilisateur avec les autres types d'expériences. Par exemple, l' « *expérience de marque* » (« *Brand experience* »), n'inclut pas seulement les interactions avec les produits de la marque, mais également les interactions avec la compagnie et ses nombreux services. De ce fait, l'expérience de marque est un concept plus large que celui de l'expérience utilisateur. Cette *expérience de marque* influence grandement l'expérience utilisateur lors de l'interaction. En effet, on pardonnera facilement les défauts d'une marque que l'on aime alors que l'on blâmera vivement les défauts d'une mauvaise marque (Law et al., 2009). La notion d' « *expérience produit* » (« *Product experience* »), utilisée par Desmet & Hekkert (2007), est au contraire plus étroite que celle de l'expérience utilisateur, dans le sens où tous les objets ne sont pas des produits commerciaux (ils peuvent être « *fait-main* » aussi) et

[29] http://www.allaboutux.org/

que de moins en moins de produits fonctionnent en isolation (ils fonctionnent plutôt en conjonction avec de multiples systèmes externes). L' « *expérience de service* » (« *Service experience* ») est une notion délicate à rapprocher de l'expérience utilisateur à cause de la diversité de services existants (restaurants, services publiques, sites de jeux en ligne, etc.). Les services « *de personne à personne* », n'ont rien à voir avec l'expérience utilisateur car il n'y a pas d'interface à « *utiliser* ». C'est seulement quand le service passe par une interface utilisateur que l'on peut parler d'expérience utilisateur. Ainsi, Law et al. (2009) recommandèrent l'utilisation de la notion d'expérience utilisateur dans le cadre de « ***produits, systèmes, services, et objets*** *avec lequel un individu peut interagir au travers d'une* ***interface utilisateur*** » (figure 37).

Durant la période de l'enquête, la série des standards ISO 9241 était en révision, et l'une de ses tâches a été d'esquisser une définition de l'UX. Quelques-uns des membres assignés à cette fonction participèrent à la session SIG de CHI 2008 « *Towards a shared definition of user experience* » (Law et al., 2008) afin de discuter de la manière dont les résultats de l'enquête future pourraient contribuer à affiner le projet de définition. Il n'est pas étonnant donc que la définition définitive retenue par les auteurs de la norme ISO 9241-210:2010 s'inspira fortement des retours de l'enquête. Ces derniers définissent l'expérience utilisateur comme « *les perceptions et réactions d'une personne qui résultent de l'utilisation effective et/ou anticipée d'un produit, système ou service* ». Ils la complètent par deux notes importantes. (i) « *L'expérience de l'utilisateur inclut toutes les émotions, convictions, préférences, perceptions, réactions physiques et psychologiques, comportements et réalisations de ce dernier, qui interviennent avant, pendant et après l'utilisation* » ; (ii) *« L'expérience de l'utilisateur est une conséquence de l'image de marque, la présentation, la fonctionnalité, les performances, le comportement interactif et les capacités d'assistance du système interactif, de l'état intérieur et physique de l'utilisateur résultant d'expériences passées, de ses attitudes, de ses compétences et de sa personnalité ainsi que du contexte d'utilisation* » (ISO, 2010).

Cette norme, qui révisa l'ISO 13407:1999, prolonga également la conception centrée utilisateur par l'ajout et la modulation de six principes clefs, inspirés du paradigme de l'expérience utilisateur : (i) la conception est fondée sur une compréhension explicite des utilisateurs, des tâches et des environnements, (ii) les utilisateurs sont impliqués dans la conception et le développement, (iii) la conception est dirigée et précisée par l'évaluation centrée sur l'utilisateur, (iv) le processus est itératif, (v) la conception couvre l'expérience de l'utilisateur dans son intégralité, et (vi) l'équipe de conception inclue des compétences et des points de vue pluridisciplinaires.

Enfin, une réplication du travail de Law et al. (2009), réalisée sur 758 professionnels et chercheurs de plus de 35 nationalités, parachève le travail toujours actuel de la définition de l'UX (Lallemand, Gronier, & Koenig, 2015) . Les résultats confirment ceux de la précédente enquête et des aspects du livre blanc sur l'UX de Roto et al., (2011) qui établit que l'UX serait propre à un individu, influencée par les expériences antérieures et les attentes de ce dernier. Les auteurs notent toutefois un certain clivage entre milieu académique et monde de l'entreprise d'une part, ainsi qu'entre pays anglo-saxons et francophones, qu'ils expliquent par des différences de niveau de maturité de l'UX et de son usage.

Le développement des méthodes UX

En parallèle du développement théorique de l'UX, un développement méthodologique, associé aux pratiques de conception et d'évaluation dans le domaine, émergea à partir des années 2000. Cela commença par leurs dénombrements. Un des premiers à dresser une liste de méthodes pour concevoir des produits désirables est Jordan (2000). Elle fut composée de méthodes aidant à l'inspiration lors de la conception, de méthodes d'évaluation des aspects plaisants d'un produit, et d'exemples d'approches multi-méthodes. En 2004, deux initiatives européennes, ENGAGE[30] et HUMAINE[31] (Human-Machine Interaction Network on Emotion) virent le jour pour appuyer le champ de l'ingénierie affective, via une capitalisation des connaissances et des méthodes sur le domaine. Le projet ENGAGE collecta dans ce contexte un certain nombre de méthodes et d'outils, entre 2004 et 2006, pour la conception et l'évaluation de ces systèmes (ENGAGE, 2006). Ces derniers les classèrent en deux catégories : générative et évaluative. Les méthodes évaluatives sont à leur tour divisées en trois groupes en fonction de la nature de leurs mesures : évaluation des qualités sensorielles (qualité perceptive, esthétisme, …), évaluation de l'expression/signifiance d'un produit (par scénario, expressions qualitatives), et évaluation des réactions émotionnelles (expressions faciales, réactions physiologiques,…). Le projet HUMAINE, de son côté, collecta et développa toute une série de méthodes de conception et d'évaluation, centrées sur la mesure des émotions, pour l'informatique affective. Elle englobèrent des méthodes classiques, comme des questionnaires ou des techniques d'entretien, et des outils plus spécialisés, comme des modèles informatiques de traitement de signaux physiologiques, gestuelles ou faciaux (HUMAINE, 2008a, 2008b).

À partir de 2008, toute une série d'initiatives furent entreprises pour constituer un état de l'art des méthodes disponibles dans la communauté IHM. Elles furent collectées lors de workshops, organisés durant CHI 2008, INTERACT 2009 et DPPI 2009 (Roto & Hoonhout, 2009; Roto & Väänänen-vainio-mattila, 2009; Väänänen-Vainio-Mattila, Roto, & Hassenzahl, 2008), un SIG (« *Special Interest Group* ») de CHI 2009 (Marianna Obrist, Roto, & Väänänen-Vainio-Mattila, 2009) et à partir de diverses enquêtes en ligne et états de la littérature (Vermeeren et al., 2010). En centralisant toutes ces méthodes, Roto et divers collaborateurs diffusèrent une liste de plus de 82 méthodes d'évaluation de l'UX (Tableau 7a et 7b) sur leur site internet[32]. De plus, ils créèrent une série de cours lors d'INTERACT 2011, CHI 2012 et CHI 2013 (Roto, Vermeeren, Väänänen-Vainio-Mattila, & Law, 2011; Roto, Vermeeren, Väänänen-Vainio-Mattila, Law, & Obrist, 2012, 2013) afin de les diffuser le plus largement possible dans la communauté. En se basant sur les caractéristiques de chaque méthode, Roto et al. (2011) les catégorisèrent selon différents critères : les tableaux 5 et 6 montrent des exemples de méthodes d'évaluation UX en fonction du moment d'étude de l'expérience utilisateur et de la phase de développement du produit.

30 http://www.designandemotion.org/society/engage/

31 http://emotion-research.net/

32 http://www.allaboutux.org/

Tableau 5 - Exemples de méthode d'évaluation selon différents moments d'étude de l'UX (*Roto et al., 2011*)

EVALUATION DES EMOTIONS	EVALUATION D'UN EPISODE UX	ÉVALUATION DE L'UX A LONG TERME
Observation	**Observation**	**Auto-rapportée**
Expressions vocales, faciales et posturales	Verbalisation à voix haute d'expérience	QCM, Laddering, iScale, RGT (Repertory Grid Technique)
Mesures psychophysiologique	**Auto-rapportée**	
Réactions sudatoires, cardiovasculaires, pupillaires et musculaires	Experience Sampling, AttrakDiff, entretiens, Reconstruction journalière	
Auto-rapportée		
Verbale : PANAS, AffectGrid Non-verbale : EmotionSlider, EmoCard, PrEmo		

Tableau 6 – Exemple de méthode d'évaluation selon différentes phase de développement produit (*Roto et al., 2011*)

CONCEPTION	PROTOTYPE NON-FONCTIONNEL	PROTOTYPE FONCTIONNEL
Conception visuelle	**Conception visuelle**	**Test en laboratoire**
Réactions émotionnelles	Réactions émotionnelles	Réactions émotionnelles, AttrakDiff
Description d'idées	**Interaction**	**Test sur le terrain**
Evaluation experte, Jeu de rôle, inspection basée sur la perspective	Verbalisation à voix haute d'expériences	Experience Sampling, Reconstruction journalière
		Feedback utilisateur
		QCM, Courbe UX, iScale

En parallèle à la démarche de dénombrement des méthodes, des recherches portant un regard critique sur ces dernières −en examinant leurs qualités, leurs caractéristiques et leurs utilisations− ont également été menées. Dans cet ordre d'idées, Vermeeren et al. (2010) entreprirent d'examiner l'état actuel des méthodes d'évaluation UX, afin de connaître leurs limites et de proposer des pistes d'amélioration. À partir d'un état de la littérature, d'une enquête en ligne et de l'organisation de plusieurs workshops et SIG (Obrist et al., 2009; Roto & Väänänen-Vainio-Mattila, 2009; Väänänen-Vainio-Mattila et al., 2008), ils finirent par recenser dans le domaine plus de 96 méthodes UX distinctes. Leur étude mena à trois grandes constatations, qui sont (i) une **sous-représentation** de méthodes pour certains domaines clés du domaine IHM actuel, (ii) de nombreuses **faiblesses pratiques**, et (ii) une **qualité scientifique générale fragile** :

- **(i) La sous-représentation de certains types de méthodes :** la première constatation réalisée fut que peu de méthodes étaient disponibles pour évaluer les premières phases de développement d'un produit. En effet, il n'y avait que de peu méthodes pour évaluer l'UX avant l'usage du produit alors que la dimension d'anticipation de l'usage est formellement comprise dans la définition ISO de l'UX (ISO, 2008). De même, pour l'évaluation d'un concept avant son développement, seule une méthode de l'échantillon −dite de « *l'immersion* »− demandait spécifiquement à l'évaluateur d'imaginer comment l'expérience avec le produit se déroule au fil des jours. Le nombre de méthodes permettant l'analyse de l'expérience utilisateur dans le cadre des tâches sociales ou collaboratives était également relativement faible. Cela constitue un problème d'autant plus critique que l'informatique devient de plus en plus sociale.
- **(ii) Les faiblesses pratiques :** leurs analyses montrèrent aussi que les méthodes UX manquaient de qualités pratiques, risquant ainsi de limiter leur diffusion dans le contexte

industriel. En effet, 13 méthodes sur les 96 examinées étaient des METHODES EXPERTES et seulement sept d'entre elles pouvaient se passer d'un recrutement utilisateur. Or, les METHODES EXPERTES furent créées pour leurs qualités pratiques car elles ne demandent que peu de ressources. La faible proportion de ces méthodes s'explique néanmoins par la nature subjective du paradigme UX (Hassenzahl, 2006), et par la relative jeunesse du domaine, qui n'a pas encore le capital empirique suffisant (ni même d'expert) pour la mise au point de méthodes plus heuristiques. Autre constat, plus de la moitié des méthodes ne s'appuient pas sur des mesures prédéfinies. En effet, de nombreux acteurs UX revendiquent l'utilisation d'EVALUATIONS OUVERTES et QUALITATIVES, car ils considèrent que l'utilisation de ces métriques ne dévoilerait qu'une faible partie de l'expérience utilisateur globale. Néanmoins, même si l'utilisation d'EVALUATIONS OUVERTES offre des avantages indéniables dans les premières phases d'élaboration d'un produit, leurs utilisations demandent un savoir-faire très spécialisé que toutes les entreprises ne peuvent se permettre d'avoir. Ainsi, l'utilisation de mesures validées et rapides à utiliser, pour toute une série de dimensions de l'UX, est un avantage pratique non négligeable. Enfin, nombre de ces méthodes gagneraient à être transposées en ligne. En effet, la capture de l'expérience utilisateur en ligne à l'avantage d'être rapide, peu coûteuse et « *écologique* », c'est-à-dire permettant d'étudier l'utilisateur dans un contexte réaliste, ce qui augmente la validité de l'enquête. Néanmoins, peu de méthodes de ce genre existent actuellement et les données récoltées sont souvent non structurées et longues à analyser. Un travail de rationalisation de ce processus serait donc extrêmement bénéfique pour faciliter leurs diffusions.

- Troisièmement, les analyses de Vermeeren et al. (2010) montrent que la **qualité scientifique générale de ces méthodes est fragile**. Premièrement, de nombreuses méthodes UX basent la collecte de leurs données sur des questionnaires non validés, ou alors de manière discutable. Il y a donc une nécessité de validation de ces mesures. Deuxièmement, la majeure partie de ces méthodes a été élaborée alors que la théorisation de l'UX était encore immature. De ce fait, la mise au point de modèles théoriques plus solides pourrait fortement consolider la validité présente des méthodes et mesures développées. Enfin, même si les bénéfices de la **combinaison de méthodes** UX sont reconnus par tous car caractérisée par une vision plus riche de l'expérience utilisateur et d'une plus grande qualité scientifique des données récoltées, peu de méthodologies émergent dans cette perspective. Pour les auteurs, la cause principale résulte d'une demande plus élevée de ressources et en expertises nécessaire, couplée à une difficulté de consolidation des données de nature différente. De ce fait, des recherches supplémentaires sont nécessaires pour fournir des directions aux praticiens sur les méthodes et les types de données UX à utiliser ensemble, ainsi que des procédures fiables pour les analyser de concert.

Bargas-Avila et Hornbæk (2011) complétèrent l'analyse des méthodes UX en se penchant cette fois ci sur la manière dont ces dernières sont utilisées dans la recherche empirique. En se basant sur une revue des méthodes de 66 études UX publiée entre 2005 et 2009, ils complétèrent et confirmèrent certaines constatations de Vermeeren et al. (2010) sur les voies à suivre pour améliorer les méthodes d'évaluation actuelles et leurs utilisations. Ils constatèrent ainsi (i) certaines faiblesses concernant la prise en compte de **caractéristiques clefs de l'UX**, tout comme (ii) un **manque de validité** de certains outils et procédures de recherche utilisées. Ainsi,

(i) le contexte d'utilisation et l'expérience utilisateur de l'usage anticipé (deux **facteurs clefs** de l'expérience utilisateur, souvent cités) sont rarement développés dans les papiers étudiés. En effet, de la plupart des études examinées sont dépourvues de descriptions riches du contexte. Dans un peu moins de la moitié des études (45%, n = 30), le contexte d'utilisation est pauvre (par choix épistémologique) car les chercheurs le contrôlent en conduisant leur enquête dans un environnement fixé (souvent en laboratoire). Dans un autre tiers des études (33%, n = 22), le contexte n'est pas contrôlé et aucune information n'est fournie sur l'environnement social et physique lors de l'utilisation (on constate cela pour la plupart des études menant leur enquête en ligne). Seulement 21% (n=14) des études, se basant sur des méthodes fortement qualitatives (ENQUETE CONTEXTUELLE, DESIGN PROBES), précisent certains éléments de contexte lors de l'interaction, mais de manière peu satisfaisante pour des méthodes de telle nature. L'étude de l'expérience au cours du temps est également peu représentée. Il n'y a pas de vraie étude longitudinale de l'expérience et seulement 20% des études s'intéressent à l'UX avant l'interaction (on rejoint ici le constat de Vermeeren et al., 2010). (ii) Parmi les **faiblesses méthodologiques générales**, un constat similaire est effectué à propos de la validité scientifique des méthodes utilisées et l'absence de leur combinaison, malgré une complémentarité reconnue. Par exemple, Avila et Hornbæk (2011) constatent que dans les études utilisant des questionnaires, 91% sont faits de toute pièce (seul ou en complément de questionnaires préexistants), même si un grand nombre d'entre eux −parfois très bien validés− existent déjà dans la littérature. De plus, les auteurs fournissent les items de ces questionnaires dans moins d'un tiers de ces publications (56% des études ne les fournissent pas du tout, et 12% partiellement). Une analyse plus récente (Law, van Schaik, & Roto, 2014) sur l'utilisation des questionnaires dans la littérature UX de 2010 à 2012 (n = 58), montre également que dans 28% des cas (n = 16) aucune information sur leurs qualités psychométriques n'est donnée. Toutes ces données confirment ainsi l'analyse de Vermeeren et al. (2010). Bargas-Avila et Hornbæk (2011) relevèrent également certains problèmes pour des méthodes émergentes et peu validées, dites CONSTRUCTIVES ou PROJECTIVES (sonde, collage/dessins, photographies, …). Ces méthodes, même utilisées pour stimuler l'inspiration des concepteurs, possèdent des ambiguités en terme de procédures et d'interprétation des résultats qui limitent fortement leur validité. Enfin, tout comme Vermeeren et al. (2010), Bargas-Avila et Hornbæk (2011) constatent un manque de triangulation des méthodes, qui, en insistant trop sur leurs assises méthodologiques respectives, finissent par nuire à la qualité de la recherche. Ainsi, beaucoup d'études qualitatives, négligent de rapporter les protocoles d'entretien, décrivent rarement leurs procédures d'analyse de données, se concentrent principalement sur une expérience utilisateur « générique » et contribuent à une explosion de dimensions associées, souvent redondantes. De même, de nombreuses études quantitatives se contentent de situations bien peu écologiques, dans des contextes d'utilisation trop simplifiés, d'une durée d'intéraction bien trop courte et en ne sollicitant pas toujours l'expérience auto-raportée des utilisateurs.

Les caractéristiques de la liste des méthodes UX, fournie par Roto et al. sur leur site *allaboutux.org (tableau 7a, b),* nous permettent également de compléter l'analyse de l'état des méthodes UX. Si le paradigme de l'utilisabilité a puisé massivement dans les modèles et les méthodes de la psychologie cognitive, nous constatons ici que c'est la psychologie positive et des émotions qui ont constitué, pour un quart des méthodes UX, la source académique de la

liste de Roto et al. Néanmoins, la provenance générale de ces méthodes reste extrêmement variée : psychologie (générale, sociale, développementale, …), design, marketing, ingénierie, ludologie, science de l'information, etc. De plus, on note que seulement un peu plus d'un quart des méthodes ont été créées dans le champ même de l'expérience utilisateur. De ce fait, la grande majorité des méthodes UX, importées de domaines plus ou moins lointains, est rentrée dans le champ de l'expérience utilisateur, avec plus ou moins d'adaptation. Certaines méthodes ont été reprises telles quelles, d'autres n'ont été que légèrement ajustées, et seulement une minorité profondément repensées. On peut alors s'interroger sur leur valeur écologique. Un travail rigoureux d'adaptation et de revalidation, prenant en compte les spécificités du domaine, serait à effectuer. D'autre part, nous constatons que pour les méthodes présentant une validation, ces dernières varient grandement en qualité. En effet, quand certaines ont été rigoureusement validées, par des protocoles précis et adaptés, d'autres ne l'ont pas été du tout, ou alors seulement par une illustration de cas d'utilisation. Cette disparité s'explique par la grande divergence épistémique des disciplines associées et la relative jeunesse du paradigme UX.

En reprenant le dilemme fidélité-bande passante de Cronbach (1960), nous voyons ainsi que le développement méthodologique dans le domaine de l'UX s'est grandement déplacé du côté de la « bande passante », c'est-à-dire de la variété des approches et de la richesse des points de vue. Le choix s'est porté sur l'abondance plutôt que sur la qualité. Il est crucial maintenant de « digérer » tous ces apports, en contextualisant les méthodes, en améliorant leur validité intrinsèque ou encore en les harmonisant entre elles (élimination des doublons, classification, formalisation du langage, …). Nous constatons également qu'aucune clef actuelle n'est offerte pour leurs recoupements, ce qui permettrait d'aller vers plus de « *fidélité* », c'est-à-dire une meilleure validité méthodologique. Pourtant, le croisement des approches, pour augmenter la validité des études, a été l'un des principaux arguments pour justifier l'incorporation dans le domaine d'autres approches non-conventionnelles. En effet, dans l'article d'Olson & Moran (1998), de nombreux intervenants défendirent l'utilisation d'alternatives au paradigme classique de validation expérimental par cet argument. McClelland parla ainsi du principe de triangulation, « *permettant de baser son jugement sur différentes sources via un système de renforcement et de correspondance* » (Commentaire 4, p. 287), John de « *l'utilisation de données de type multiple, sur un mode triangulatoire, pour converger vers une explication* » (Commentaire 5, p. 292), Monk de *« combinaison* » d'approche pour élargir la portée des leçons tirées de la recherche (Commentaire 6, p. 302), Oviatt de « *la combinaison de méthodes* », car *« la diversité méthodologique est notre plus grande richesse et l'utilisation de multiples méthodes convergentes notre technique de conception la plus puissante* « (Commentaire 7, p. 304), Carroll de « *complémentarité entre les approches* » (Commentaire 8, p. 310) et enfin Mackay de souligner la valeur de la multiplicité des méthodes d'évaluation, triangulées à l'intérieur et entre les disciplines, pour fournir une vision d'ensemble la plus large possible (Commentaire 9, p. 312-315). Pourtant, derrière toutes ces déclarations, on constate aujourd'hui que très peu de choses ont été faites pour communier ces différentes approches. Au mieux, l'apport de toutes ces nouvelles méthodologies a déplacé la guerre des chapelles paradigmatiques de l'extérieur à l'intérieur du domaine. Des études permettant de dépasser réellement leurs simples confrontations restent encore à imaginer.

Tableau 7a - Liste chronologique (#-M) des méthodes d'évaluation dans le domaine des systèmes collaboratifs

Méthodes de conception et d'évaluation UX	Sources de la description et/ou de la validation
2DES (Two Dimentional Emotion Space)	Schubert (1999, 2001)
3E (Expressing Experiences and Emotions)	Tähti & Arhippainen (2004)
Aesthetics scale	Laviea & Tractinsky (2004)
Affect Grid	Russell, Weiss, & Mendelsohn (1989)
Affective Diary	Ståhl, Höök, Svensson, Taylor, & Combetto, (2009)
Attrak-Work questionnaire	Väätäjä, Koponen, & Roto (2009)
AttrakDiff	Hassenzahl, Burmester, & Koller (2003)
AXE (Anticipated eXperience Evaluation)	Gegner & Runonen (2012)
Co-discovery	Jordan (2002)
Context-aware ESM	Froehlich et al. (2007); Intille et al. (2003)
Contextual Laddering	Grunert & Bech-Larsen (2005); Gutman (1982); Vanden Abeele, Zaman, & De Grooff (2012)
Controlled observation	Jordan (2002)
Day Reconstruction Method	Karapanos et al. (2009)
DES (Differential Emotions Scale)	Izard (1972)
EDA (ElectroDermal Activity)	Ward & Marsden (2003); Jung (1907); Dawson, Schell, & Filion (2007)
Emo2	Laurans & Desmet (2006)
Emocards	Desmet, Overbeeke, & Tax (2001)
Emofaces	Bradley & Lang (1994); Desmet et al. (2001)
Emoscope	Bustillo (2007)
ESD (Emotion Sampling Device)	Roseman, Antoniou, & Jose (1996)
Experience clip	Isomursu, Kuutti, & Väinämö (2004)
ESM (Experience Sampling Method)	Csikszentmihalyi & Larson (1992) ; Scollon, Kim-Prieto, & Diener, (2003)
Experiential Contextual Inquiry	Beyer & Holtzblatt (1998)
Exploration test	Kuniavsky (2003); Young & Veen (2008)
Facereader	den Uyl, van Kuilenburg, & Lebert (2005)
Facial EMG	Dimberg (1990); Mandryk & Atkins (2007)
Feeltrace	Cowie et al. (2000)
Fun Toolkit	Read & MacFarlane (2006)
GEQ (Game Experience Questionnaire)	Ijsselsteijn, Poels, & de Kort (2008)
Geneva Appraisal Questionnaire	Scherer (2001)
Geneva Emotion Wheel	Scherer (2005)
Group-based expert walkthrough	Foelstad (2007)
HED/UT (Hedonic Utility scale)	Voss, Spangenberg, & Grohmann (2003)
Human Computer trust	Madsen & Gregor (2000)
I.D. Tool	Opperud (2004)
Immersion	Jordan (2002)
IMI (Intrinsic Motivation Inventory)	Ryan (1982)
iScale	Karapanos, Martens, & Hassenzahl (2009, 2010)
Kansei Engineering Software	Schütte (2006)
Living Lab Method	Abowd et al. (2000); Niitamo, Kulkki, Eriksson, & Hribernik (2006)
Long term diary study	Bolger, Davis, & Rafaeli (2003)
Mental effort	Meijman, Zijlstra, Kompier, & Mulders (1986)
Mental mapping	Hine (1995); Jordan (2002)

Tableau 7b - Liste chronologique (O-W) des méthodes d'évaluation dans le domaine des systèmes collaboratifs

Méthodes de conception et d'évaluation UX	Sources de la description et/ou de la validation
OPOS (Outdoor Play Observation Scheme)	Bakker, Markopoulos, & de Kort (2008)
PAD (Pleasure Arousal Dominance)	Mehrabian & de Wetter (1987)
Paired comparison	Thurstone (1927)
Perceived Comfort Assessment	Helander & Zhang (1997); Helander (2003)
OPOS (Outdoor Play Observation Scheme)	Bakker, Markopoulos, & de Kort (2008)
PAD (Pleasure Arousal Dominance)	Mehrabian & de Wetter (1987)
Perspective-Based Inspection	Zhang, Basili, & Shneiderman (1999)
QSA GQM questionnaires	Pellerey (1996); Solingen & Berghout (1999)
Reaction checklists	Jordan (2002)
RGT (Repertory Grid Technique)	Hassenzahl & Wessler (2000); Kelly (1955); Bannister & Fransella (1985)
SAM (Self Assessment Scale)	Lang (1980); Fang & Sun (2014); Serrano, Botella, Baños, & Alcañiz (2013)
Semi-structured experience interview	Silverman (2005); Mason (1996)
Sensual Evaluation Instrument	Isbister, Hook, Laaksolahti, & Sharp (2007);
Sentence Completion	Kujala & Nurkka (2009, 2012); Kujala, Walsh, Nurkka, & Crisan (2013)
ServUX questionnaire	Väänänen-Vainio-Mattila & Segerståhl (2009); Väänänen-Vainio-Mattila, Väätäjä, & Vainio (2009)
SUMI (Software Usability Measurement Inventory)	Kirakowski & Corbett (1993)
This-or-that	Zaman & Abeele (2007); Sim & Horton (2012)
Timed ESM	Hektner, Schmidt, & Csikszentmihalyi (2007)
TRUE (Tracking Realtime User Experience)	Kim et al. (2008)
TUMCAT (Testbed for User experience (UX) measurement of Mobile, Context-Aware applications)	Vermeeren, Kort, Cremers, Smets, & Fokker (2008); Kort, Vermeeren, & Fokker (2007)
UTAUT (Unified Theory of Acceptance and Use of Technology	Venkatesh, Morris, Davis, & Davis (2003)
UX Curve	Kujala et al. (2011)
UX Expert evaluation	Väänänen-Vainio-Mattila & Wäljas (2009, 2010)
UX laddering	Reynolds & Gutman (1988) ; Abeele et al. (2012)
Valence method	Burmester, Mast, Jäger, & Homans (2010)
WAMMI (Website Analysis and Measurement Inventory)	Kirakowski & Cierlik (1998)
Workshops + probe interviews	Lucero (2009)

Enfin, il convient de préciser que si des recherches méthodologiques visant à combler toutes ces lacunes sont encore rares et balbutiantes, cela l'est pour plusieurs raisons. Premièrement, l'engouement académique vers le qualitatif, par l'entrée de nouveaux acteurs venant de secteurs plus artistiques, a freiné d'un point de vue épistémologique et idéologique l'intérêt que s'est fait la communauté sur de telles études (Law & van Schaik, 2010). De plus, le domaine d'étude étant encore jeune, ce dernier point demande encore une compréhension plus fine de l'UX afin de construire, à partir de celle-ci, des méthodes plus valides (Law, Bevan, et al., 2008).

Dans le sillage de l'UX : l'innovation et le développement de l'ergonomie prospective

Depuis ces dernières années, une autre tendance, entraînée par la mutation des modèles économiques et la capacité des technologies, a incité le domaine de l'ergonomie à incorporer dans ses méthodes l'innovation et la créativité. Ce nouvel enjeu, vital dans l'économie d'aujourd'hui, a obligé le domaine de l'évaluation des systèmes numériques à compléter, prolonger et adapter ses méthodologies. Cette révolution, en amont du cycle classique de conception, complète avantageusement le paradigme UX.

La mutation de l'économie et le marché de l'innovation

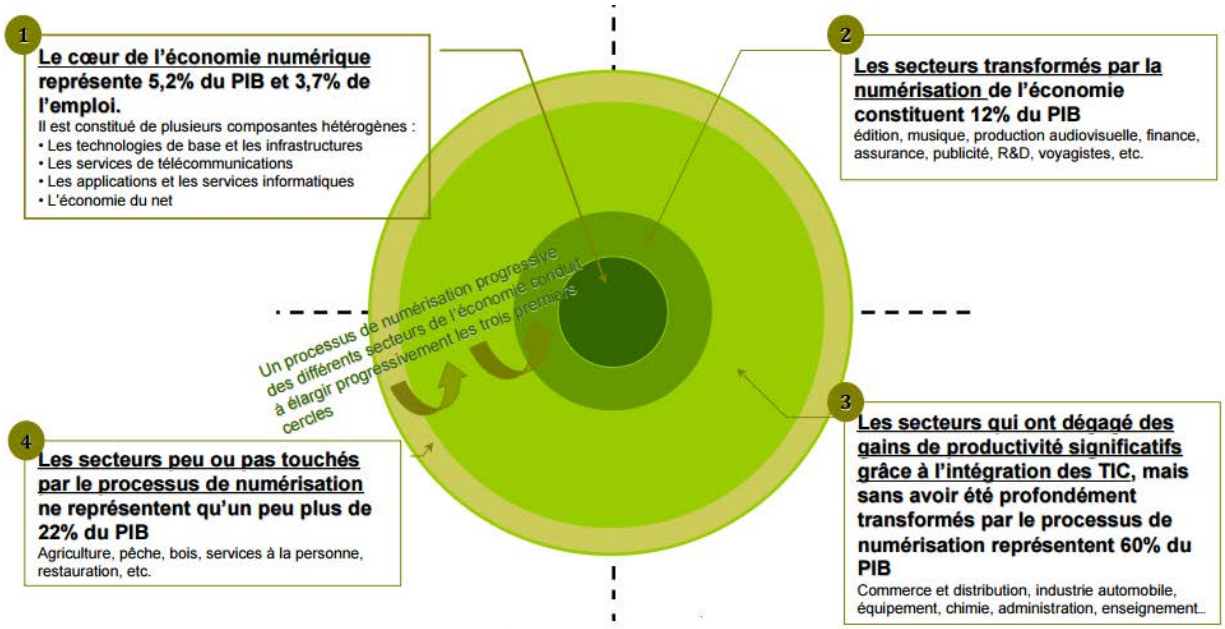


Figure 38 – La part de l'économie française concerné par l'économie numérique (tirée de Siné et al., 2012)

Le numérique joue un rôle de plus en plus important dans l'économie et l'emploi (figure 38). En effet, si le cœur de l'économie numérique représente 5,2% du PIB et 3,7 millions d'emplois, c'est près de 80% de l'économie française qui en est liée (Siné, Hausswalt, & Garcin, 2012). En augmentant la compétitivité de tous les secteurs de l'économie, les nouvelles technologies de l'information et de la communication ont contribuées à plus de 50% de la croissance de la productivité en Europe, de 2000 à 2004[33]. En fort développement, ses perspectives de croissance sont considérables, avec plus de 8% de croissance par an pour l'économie d'Internet d'ici 2016 (Dean et al., 2012).

[33] Communication de la Commission au Conseil, au Parlement européen, Comité économique et social européen et au Comité des régions - «i2010 – Rapport annuel 2007 sur la société de l'information», Bruxelles, 30 Mars 2007 (disponible sur http://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:52007DC0146&from=FR)

L'économie numérique accélère aussi de facto le rythme de l'innovation et de la diffusion des nouveaux biens et services (DeGusta, 2012). En effet, s'il a fallu des décennies pour équiper 50% des foyers américains du téléphone, il a fallu moins de cinq ans pour faire de même pour les smartphones (Figure 39). Pour les applications en ligne, cela a été encore plus rapide. Mis en ligne en 2004, Facebook est passé de deux cents millions d'utilisateurs en 2009, à huit cents millions en 2011, et à un milliard un an plus tard. La traction sur Internet la plus rapide à ce jour est détenue par l'application *Pinterest*, qui a atteint le cap des dix millions de visiteurs uniques par mois, sur une période d'un an, de février 2011 à Janvier 2012[34]. On notera que d'autres secteurs d'activité, comme celui de l'automobile, ont également profité de cette accélération, en passant d'un cycle de vie produit de 60 mois à des cycles de 24-36 mois, en moins de cinq ans.

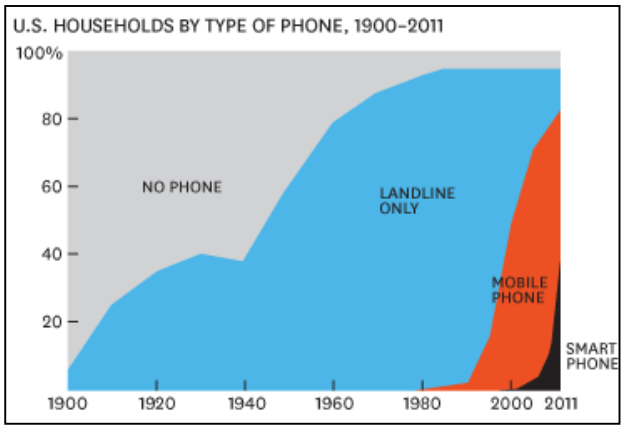


Figure 39 – De sans téléphone aux smartphones (tiré de DeGusta *(2012)*

La qualité des outils numériques de conception, mis à profit dans l'innovation, y a joué un rôle important. Mais la nature du numérique lui-même a également conduit à cette accélération. Auparavant, avec les technologies classiques, il s'agissait de transformer lentement la nature et la matière pour les soumettre aux exigences humaines. Le numérique, se basant sur une machine indéterminée, d'un microprocesseur d'une certaine manière toujours inachevé, possède en lui la flexibilité qui permet d'atteindre une vitesse renouvellement des produits et des services sans commune mesure (Brangier & Robert, 2013). Le Web 1.0 puis le Web 2.0 ont été autant de facteurs d'accélération supplémentaires pour augmenter le rythme d'innovation et d'adoption des nouveaux usages de manière exponentielle. Le Web 1.0, en s'appuyant sur une diffusion sans frontière et à un coût négligeable, a permis l'avènement de la start-up internet. Cette dernière se distingue d'une PME innovante par le fait qu'elle est conçue dès le départ pour grandir fortement et rapidement : elle vise dès l'amorçage une « *scalabilité* » qui lui permet de croître de façon exponentielle, en s'appuyant sur les avantages uniques que lui fournit l'infrastructure numérique globalisée. De plus, accélérant encore plus la diffusion, les réseaux sociaux, en 2007, ont permis l'émergence d'une nouvelle forme de diffusion de masse, que l'on peut aisément qualifier de virale. Fogg (2008), spécialiste de la persuasion informatique, l'appelle la « *persuasion interpersonnelle de masse* » (« *Mass Interpersonal Persuasion* » ou MIP). Cette nouvelle forme de persuasion explique la vitesse d'adoption des applications en ligne d'aujourd'hui. La MIP est composée de six éléments, qui, avant Facebook, n'avaient jamais été tous réunis ensemble : une expérience persuasive, une structure automatisée, une distribution sociale, des cycles rapides, un graphe social immense et un système de mesure de l'impact (Fogg, 2008).

[34] Josh CONSTINE (7 février 2012). Pinterest Hits 10 Million U.S. Monthly Uniques Faster Than Any Standalone Site Ever – comScore, Techcrunch, (disponible sur http://techcrunch.com/2012/02/07/pinterest-monthly-uniques/

Une autre caractéristique de l'économie numérique est sa tendance à conduire fréquemment ses acteurs à l'acquisition de positions dominantes. En effet, il existe peu de cas d'entreprises venant en concurrencer une autre sur un produit comparable ; et cela même si ces entreprises sont très puissantes, comme nous illustrent l'échec et le semi-échec de Google+ et Bing. Cette situation de « *verrouillage* » (Arthur, 1989) provient de plusieurs facteurs. Un premier facteur, déjà développé, est l'effet de réseau, qui correspond au fait que, plus il y a de personnes qui utilisent un service, plus celui-ci prend de la valeur, et attire de ce fait encore plus d'utilisateurs (Klemperer, 2006). À partir d'une certaine taille, d'autres manœuvres plus ou moins licites peuvent être entreprises pour garder la mainmise sur son capital d'utilisateurs (Boullier, 2012) :

- arriver à une manipulation de la sélectivité des capacités du réseau : en s'associant avec des opérateurs ou en influant sur la régulation/standardisation, de façon à permettre par exemple un débit différencié selon les contenus des paquets ;
- se positionner comme plate-forme : pour capter la valeur créée à l'extérieur de l'organisation et accélérer les rendements d'échelle nécessaires à l'établissement d'une position dominante ;
- mettre au point une stratégie d'enclosure/lock-in : par des brevets ou la création d'un écosystème clos reposant sur une incompatibilité totale des données ou matériels avec les services ou solutions tierces.

Ainsi, fuyant la confrontation directe, les entreprises numériques se livrent plutôt à une concurrence caractéristique de la frontière technologique, fondée sur la différenciation. On assiste alors à des compétitions du type « *Winner takes all* » (Frank & Cook, 1995), où l'objectif est d'arriver le premier dans un domaine vierge et de s'y installer de façon la plus durable possible. C'est en général une stratégie en deux temps : une phase d'amorçage suivie d'une phase d'expansion. En phase d'amorçage, le potentiel d'innovation est décuplé par l'effort de différenciation, qui, en se positionnant dans un marché encore immature, lui permet d'arriver à une situation de position dominante dans des délais très courts. Puis, en phase d'expansion, lorsque la marque et l'infrastructure sont assez solides, il devient alors nécessaire de prendre des positions sur des marchés connexes pour garder un avantage compétitif (Collin & Colin, 2013).

Enfin, malgré la constitution de véritables oligopoles numériques, la position des différents acteurs, quels qu'ils soient, reste précaire. En effet, l'économie du numérique est en perpétuelle et rapide mutation. Aucun service, aucune technologie et aucun modèle économique ne peut être considéré comme pérenne. L'intensité d'innovation, permise par le progrès technologique et le laboratoire internet (véritable plateforme d'expérimentation), pousse l'économie du numérique à se renouveler sans cesse. Google a éclipsé toute une génération de moteurs de recherche grâce à son approche innovante de l'indexation. Apple, proche de la faillite en 1997, a remonté la pente de manière spectaculaire suite aux visions disruptives de Steve Jobs, qui en a repris la tête, et transformé des secteurs entiers de l'économie.

L'innovation perpétuelle dans ces entreprises est donc une condition moderne de leur survie. Elle peut prendre deux formes : incrémentale ou disruptive. L'innovation incrémentale, qui est une forme d'amélioration continue, est à la faveur des entreprises établies qui l'utilisent pour renforcer leur domination. L'innovation disruptive, en rupture avec l'existant, est préférable pour les nouveaux entrants, car ces derniers sont plus efficaces et moins conservateurs que les

acteurs déjà en place quand il s'agit de porter des idées très novatrices sur le marché (Christensen, 2013). L'innovation dans la conception de produits numériques est devenue donc un enjeu majeur qui se traduit par un bouleversement de l'organisation de l'entreprise et de ses méthodes de développement.

La gestion de l'innovation dans l'entreprise d'aujourd'hui

L'une des conceptions les plus anciennes de l'innovation a été de la voir comme le prolongement naturel de la recherche : « *la science invente, l'industrie applique et la société suit* »[35]. Ce processus d'innovation, « *techno-push* » ou « *science-push* », déclare que ce sont les découvertes scientifiques qui vont donner naissance aux innovations, le marché étant perçu seulement comme un réceptacle. Ce modèle linéaire a été largement critiqué, car il n'admet aucun aller-retour (Kline & Rosenberg, 1986). Nous avons vu que d'autres modèles existent, comme le modèle de conception centré utilisateurs, très itératif et au contact permanent du marché. Néanmoins, ce processus est mis en place en général seulement après que les grandes lignes du projet ont été définies. Or, il peut exister dès le début un décalage entre celui-ci et les besoins réels de la société. Ce décalage peut lui être fatal, car la confrontation avec le marché peut arriver trop tard. Pour innover donc, il faudra prendre en compte, dès la racine du projet, ses faces techniques et sociales. En effet, la plupart des innovations « *techniques* », sont en fait des innovations sociotechniques, parce que les compétences organisationnelles, les interactions sociales, les usages sont renouvelés en même temps que les objets techniques eux-mêmes (Callon, Joly, & Rip, 2010). Pour équilibrer le modèle de conception, il faut donc faire remonter, dans le pilotage de l'innovation en amont, les contraintes de l'aval (propre au marché) : « *Il n'est plus possible de raisonner sur un sujet uniquement en fonction de préoccupations scientifiques et techniques. Les chercheurs doivent prendre en compte et assimiler les dimensions industrielles et commerciales, alors que l'enjeu sur des marchés de spécialité réside dans la capacité à proposer de nouveaux produits porteurs de valeur d'usage pour les clients industriels, et au-delà pour les consommateurs finaux* » (Midler, 2003). Une invention technique, sans bénéfice dans la société, ne deviendra pas une innovation. Par contre, une innovation −utile et désirable− qui se base sur une technologie sans pareil est un avantage de taille pour une entreprise. En effet, posséder une innovation qui repose sur une ressource inimitable ou rare (brevet, complexité technique, …) est un avantage concurrentiel reconnu (Barney, 1991). Néanmoins, les technologies étant de plus en plus ouvertes et flexibles, c'est l'identification des usages futurs qui fait souvent la différence entre les entreprises aujourd'hui (Brangier & Robert, 2013). Un des autres problèmes du modèle linéaire de l'innovation, basé sur l'excellence technique, est la surfixation mentale que ces connaissance uniques entrainent pour la stratégie de l'entreprise : ses « *Core capabilities* » (Prahalad & Hamel, 1990) se transformant alors en « *Core rigidities* » (Leonard-Barton, 1992). Ainsi, l'empire de Kodak, régnant sur la photographie depuis 1880, s'est effondré dans les années 1990 car n'ayant pas su se réinventer (Sauteron, 2009). Pour Teece, Pisano et Shuen (1997), une entreprise doit au contraire développer ses « *Dynamic capabilities* » : c'est-à-dire son habilité à intégrer, construire et reconfigurer ses compétences internes et externes pour s'adapter à des environnements socio-technico-économiques turbulents. En d'autres mots, cela reflète la

[35] Selon le slogan de l'Exposition Mondiale de Chicago en 1933.

capacité d'une entreprise à se créer de nouvelles formes d'avantages compétitifs, étant donné sa dépendance au sentier[36], c'est à dire son positionnement traditionnel sur le marché.

L'entreprise innovante est donc une entreprise **apprenante** qui sait se reconfigurer quand il le faut. Elle doit s'instruire en permanence, en s'appuyant sur des systèmes de veille et de gestion des connaissances : le « *Knowledge Management* » (Alavi & Leidner, 2001). Elle peut également utiliser des « *Gatekeepers* », dont le rôle est de servir d'interface entre l'entreprise et l'extérieur, mais aussi en son sein, en étant capable de porter et d'absorber ces nouvelles connaissances (Cohen & Levinthal, 1990). Cependant, la multiplicité et la complexité des savoirs à acquérir aujourd'hui pour innover peuvent largement déborder les ressources d'une entreprise seule, quelle que soit sa détermination. C'est pourquoi, beaucoup d'initiatives de coopération ont émergé sous différents noms: « *open innovation »* (Chesbrough, Vanhaverbeke, & West, 2006; Christensen, Olesen, & Kjaer, 2005; Huizingh, 2011), « *co-innovation »* (Bossink, 2002; S. M. Lee, Olson, & Trimi, 2012; Maniak & Midler, 2008), « *partenariats d'innovation »* (Fréchet, 2002; Vanhée, 2008), « *partenariats d'exploration »* (Gillier, 2010; Segrestin, 2006), « alliance d'exploration *»* (Koza & Lewin, 1998), etc. En remontant dans le processus de conception et en se déplaçant dans les phases amont de l'innovation (Midler, Maniak, & Beaume, 2007), ces collaborations visent à découvrir de nouvelles opportunités de marché (Koza & Lewin, 1998) en défrichant collectivement des terrains dont les connaissances restent limitées (Segrestin, 2006). De plus, la rencontre d'entités aux compétences très variées est en général une entreprise fertile du point de vue de l'innovation, car celle-ci est fortement liée à l'échange et la combinaison originale de connaissances provenant de domaines généralement séparés (Hargadon & Sutton, 2001; Justesen, 2001). Cela va dans le sens de la nouvelle signature de l'université de Lorraine à l'occasion de son lancement, en 2012 : « *faire dialoguer les savoirs, c'est innover* ». Toutefois, quand les alliances réunissent des organisations dont les objectifs sont parfois fortement concurrents, des tensions peuvent naître et nuire à la collaboration. Le concept de « *coopétition »* reflète cette dualité entre intérêts particuliers et collectifs (Blanchot & Fort, 2007).

En parallèle, l'entreprise innovante doit pouvoir créer en son sein un climat propice à la créativité. En effet, si l'on définit la créativité comme la « *capacité individuelle et collective d'imaginer un concept neuf, un objet inédit, un produit novateur, une solution nouvelle ou simplement un fait original* » (Brangier & Robert, 2013, §34), la mise en place d'un environnement favorable à son expression lui est vitale. De nombreuses études ont tenté de déceler les facteurs organisationnels et les leviers managériaux permettant à une entreprise d'être créative : mélange d'expertise, encouragement de la part des supérieurs, droit à l'erreur, … (Amabile, 1996, 1998; Woodman, Sawyer, & Griffin, 1993). Il est également possible de jouer sur la motivation des employés par l' « *intrapreneurship »* (Menzel, Aaltio, & Ulijn, 2007) qui permet de valoriser les initiatives avec le soutien de la hiérarchie. Il est également bon pour les entreprises créatives de s'écarter de la rigidité procédurière des structures bureaucratiques (Mintzberg, 2004), voire pour les équipes en charge de l'innovation de s'écarter de l'entreprise

[36] Tendance à se conforter à des décisions passées, justifiées à une époque, mais qui ont cessé d'être optimales ou rationnelles, et qui perdurent parce que les changer impliquerait un coût ou un effort trop élevé à un moment, alors que ce changement pourrait être payant à long terme.

tout court : « *Keep the disruptive organization independent* » (Bower & Christensen, 1995, p. 52). En effet, l'inertie des grosses structures nuit à la réactivité et l'audace nécessaire pour innover. Boullier (2010b) s'y rapporte par la notion de « *performance par la vitesse* » : « *La dimension de la performance reste vitale, mais au lieu de la traduire en "big is beautiful", qui est un raisonnement financier, il faut la traduire en exigence de vitesse (non pas de court terme mais de réactivité, ce qui n'est pas la même chose). Ce qui veut dire agir sur le cycle de développement en admettant comme règle qu'une belle innovation qui reste dans les labos n'a aucun intérêt alors qu'une innovation imparfaite, qui prend le risque du marché et de ses utilisateurs et qui se donnent les moyens de réviser très vite ses présupposés en fonction des retours, a toutes les chances de survivre et de se coupler avec son public » (p. 10).*

Les méthodologies en ergonomie pour les technologies émergentes et les applications innovantes

Étant donné les enjeux de l'innovation dans l'économie moderne, l'ergonomie et ses méthodes ont connu ces dernières années une série d'évolutions qui lui ont permis de se positionner efficacement sur le créneau de la création de produits et de services nouveaux, même si l'ergonomie a toujours eu l'innovation pour composante, dans une part plus ou moins importante de son action. En effet, cette dernière s'est exprimée de manière grandissante en fonction du changement des demandes d'intervention, passant progressivement de la correction, puis à la conception, pour arriver enfin à la prospection. Il est intéressant de souligner que le défi représenté par les nouvelles technologies, pour l'ergonomie, a été prévu par Bartlett dès les années 60. Ce dernier prédit le rôle futur de l'ergonomie comme celui de « *devancer la technologie et de prescrire les choix technologiques en accord avec les caractéristiques et les besoins humains* » (Bartlett, 1962). Néanmoins, ce rôle ne fût jamais encore vraiment tenu et il est juste de dire que l'innovation n'a pas été historiquement l'objectif principal des ergonomes. De ce fait, la discipline s'est retrouvée jusqu'il y a peu limitée par ses méthodes. En effet, l'ergonomie s'est longtemps appuyée sur les méthodes classiques de l'analyse du travail pour prédire les usages, en analysant les besoins des utilisateurs dans des situations déjà existantes. La limite évidente de cette méthode est qu'il n'y a pas toujours de situation à observer et donc de nouveaux besoins ou usages à faire remonter dans la conception d'un nouveau produit. Dans certains cas, l'ergonome se retrouve contraint d'étudier des technologies à l'état de projet et dont personne ne sait encore quoi en faire (Brangier & Bastien, 2006). Ainsi, étant donné que l'innovation bouleverse l'activité avant de la structurer, il est clair que l'analyse de l'activité n'est pas un prédicteur suffisant des usages futurs (Brangier & Robert, 2013).

L'enjeu a été donc de développer des méthodes permettant de prédire ces usages futurs. C'est pour cet objectif qu'a été conçue l'ergonomie prospective (Robert & Brangier, 2009). Un numéro spécial de la revue « *Le travail Humain* » lui a été d'ailleurs dernièrement consacré[37]. La prospective est définie par le petit Robert par « (…) *l'ensemble des recherches concernant l'évolution future de l'humanité et permettant de dégager des éléments de prévision* ».

[37] http://www.cairn.info/revue-le-travail-humain-2014-1.htm

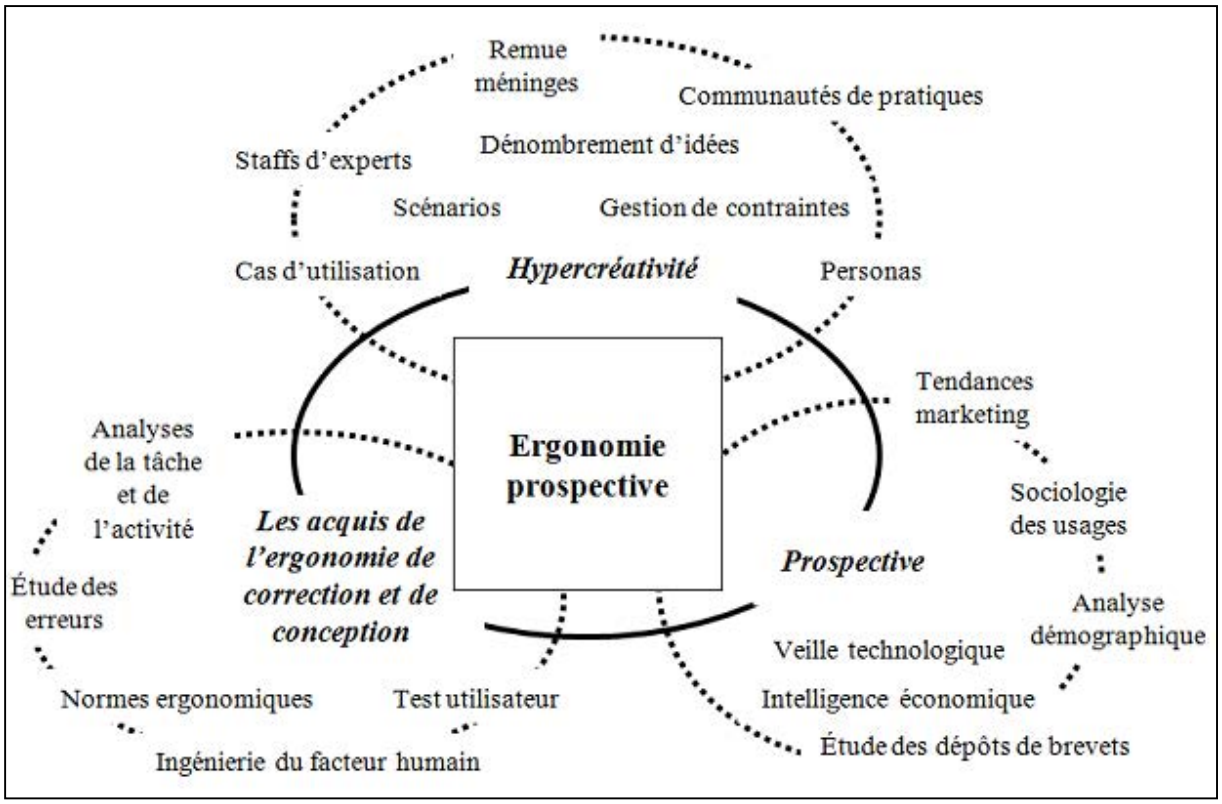


Figure 40 – Les facettes de l'ergonomie prospective (tirée de *Brangier & Robert, 2013)*

L'ergonomie prospective vise ainsi à prédire l'évolution future des usages en enrichissant les outils classiques de l'analyse ergonomique par de nouvelles méthodes et concepts. Brangier et Robert (2013) regroupent ces ajouts méthodologiques dans deux nouvelles démarches venant se greffer aux outils classiques de l'ergonomie : la démarche PROSPECTIVE et la démarche de l'HYPER-CREATIVITE (Figure 40).

La PROSPECTIVE représente une démarche relativement organisée pour donner de la lisibilité sur des futurs possibles en cernant des tendances (démographique, sociologique, technologique, économique) qui peuvent nous renseigner sur le sens pris par l'innovation (Robert, 2003). Ces outils de veille et d'intelligence économique ont pour but de détecter des « *signaux faibles* », c'est-à-dire, des informations partielles et fragmentaires fournies par l'environnement, qui, par leurs analyses et leurs recoupements, nous permettraient potentiellement d'anticiper les futurs besoins clients ou les marchés porteurs à l'avenir (Caron-Fasan, 1998; Lesca, 1996). En outre, la participation de COMMUNAUTES D'EXPERTS (Bastien et al., 2009; Brangier, Dinet, & Bastien, 2009) ou de LEAD USERS (Von Hippel, 1986) à cette réflexion est également un moyen efficace de détecter ces besoins avant leur émergence sur le marché de masse. La méthode dite de COMMUNAUTE D'EXPERT fait appel à la fois à des personnes issues de communautés d'intérêt (c'est-à-dire d'un ensemble de personnes partageant un même usage et qui représente, à ce titre, des utilisateurs ciblés par le produit à concevoir) et de communautés de connaissances (c'est-à-dire d'un ensemble de personnes qui disposent d'un ensemble de savoirs reconnus sur les usages des utilisateurs ciblés par le produit à concevoir). La méthode des LEAD USERS (Figure 41) fait appel à une classe d'utilisateurs capables de (1) détecter en avance des besoins qui deviendront bien plus tard ceux des utilisateurs de masse et (2) de proposer des solutions adéquates pour y répondre (Von Hippel, 1986).

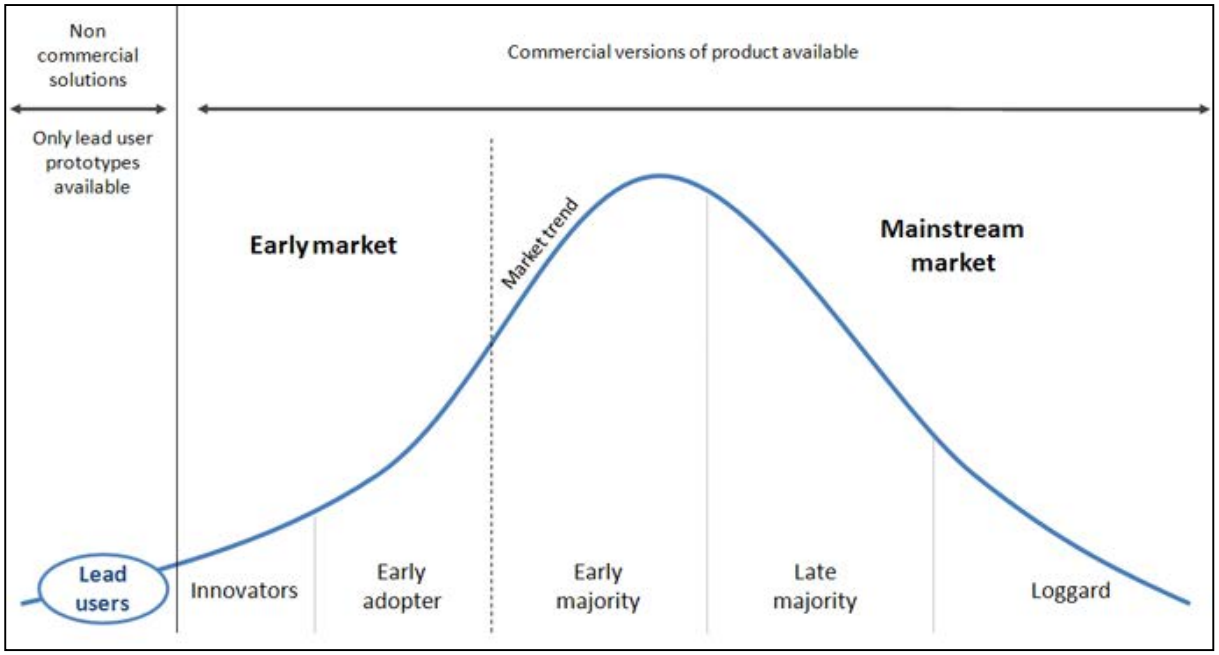


Figure 41 – Les Leads Users et les tendances du marché (tirée de strategies4innovation.wordpress.com)

L'HYPER-CREATIVITE représente une démarche visant à faciliter le processus d'idéation par l'organisation d'un bouillonnement créatif (Brangier & Robert, 2013). Elle mobilise différentes techniques visant à stimuler la production d'idées et à améliorer leurs qualités par diverses stratégies et heuristiques de raisonnement : analogies, associations, approches systématiques, psychologiques, … La technique la plus connue d'entre elles est celle du BRAINSTORMING (Osborn, 1953), laquelle consiste à stimuler la créativité des individus en s'appuyant sur des procédés d'associations et de rebonds d'idées. Dans le domaine de la créativité, de la conception (tout comme dans tout système de résolution de problème) il est possible de distinguer deux phases successives (Figure 42) : une phase de divergence, durant laquelle des idées naissent et élargissent l'espace de recherche ; puis une phase de convergence, où les solutions proposées sont mises à l'épreuve et sélectionnées.

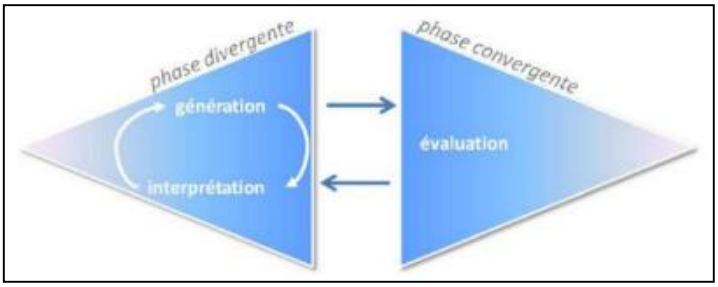


Figure 42 - Modèle de divergence – Convergence (tirée de *Guerlesquin, 2012)*

La démarche prospective et créative est placée en tout début du processus afin de préparer la conception des nouveaux produits. Cette étape est appelée « *phase d'avant-projet* » (Gautier & Lenfle, 2004) ou encore « *Fuzzy Front End* » (Reid & de Brentani, 2004). Elle a pour objectif de spécifier le « *conceptual design* », c'est à dire d'établir un nouveau concept, basé sur des choix techniques particuliers et à faire une première estimation de sa valorisation commerciale (Pahl, Wallace, & Blessing, 2007). C'est une étape cruciale et extrêmement stratégique pour le projet car les choix initiaux vont grandement affecter sa réussite future. En effet, même si les phases amont ne représentent qu'environ 20% du budget de conception, les décisions prises durant cette phase engageront 80% des dépenses totales du projet (Guerlesquin, 2012). De plus,

par leur grande liberté d'action, les phases amont sont plus propices à une intervention sur le futur des produits. Mais il y a un problème : les connaissances nécessaires pour guider la conception se développent au fur et à mesure du projet, au moment où ce dernier devient de plus en plus dur à changer. Ullman (2010) résume ce paradoxe de la manière suivante : « *Plus vous apprenez, moins vous avez la liberté d'utiliser ce que vous savez* ». La figure 43 schématise ce rapport, avec, dans les phases amont, une connaissance faible du problème et une liberté d'action en conception grande ; puis, dans les phases aval, un accroissement de la connaissance du problème, accompagné d'une diminution de la liberté de conception. Ainsi, Zeiler et al. (2007) montrent que l'influence d'une information sur le produit conçu évolue tout au long du processus, avec un impact maximal lors des phases amont du processus de conception.

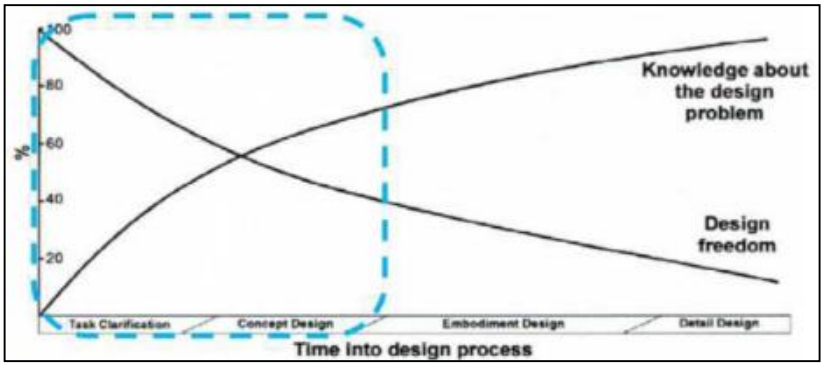


Figure 43 - rapport « connaissance du problème / liberté d'action » en fonction de l'avancée dans le processus de conception (tirée de *Guerlesquin, 2012)*

Dans le cas de l'étude de technologies émergentes, plusieurs méthodes de créativité peuvent être utilisées en fonction de la maturation conceptuelle et technique des technologies, combinées aux choix stratégique de l'entreprise. Pour Gillier et Piat (2011), chaque technologie est composée de deux dimensions étroitement connectées : une dimension technique et une dimension d'usage. La dimension technique renvoie aux propriétés phénoménologiques qui donnent à la technologie sa forme : ses propriétés physico-chimiques, ses fonctions techniques, … La dimension d'usage renvoie à la façon dont la technologie sera ou pourra être utilisée en fonction des besoins utilisateurs et du marché. Durant le stage d'émergence d'une technologie, ces deux dimensions ne sont, par définition, pas encore formellement établies : la délimitation technique est encore mal définie et les usages ciblés ne sont pas encore validés. De ce fait, la tâche créative, consistant à trouver des applications valides pour une technologie émergente, réside dans la formulation d'une paire adéquate des dimensions « technique/usage », qui conduira à la définition de l'identité de la technologie (Gillier & Piat, 2011). Plusieurs situations sont possibles et vont solliciter différentes approches méthodologiques :

- **Dimension technique stable et dimension d'usage instable** : Il s'agit du modèle d'innovation techno-push : une (/des) technologie(s), souvent nouvelle(s), est (/sont) sélectionnée(s), mais ses (/leurs) usages ne sont pas encore fixés. L'utilisation de LEAD USER est alors souvent mis en avant car ces derniers sont capables de voir les applications futures des technologies émergentes avant les autres (Urban & von Hippel, 1988) et selon des modalités plus innovantes et disruptives que les utilisateurs de masse (Magnusson, 2009; Eric von Hippel,


Figure 44 – TechCards utilisées au sein des Bell Labs

2014). De plus, quand, dans l'entreprise, le panel des technologies à valoriser est important, l'utilisation d'objet intermédiaire sous la forme de carte (Figure 44) permet de faciliter la communication dans les équipes multidisciplinaires et d'améliorer la qualité et le nombre d'idées générées lors des sessions de créativité (Ocnarescu, Rodio, Eve, Labrune, & Bouchard, 2011).

- **Dimension technique instable et dimension d'usage stable** : dans cette situation, une application initiale existe déjà mais demande à être améliorée. Dans ce cas, de nouvelles technologies sont examinées pour dépasser les performances de l'ancienne. Cette analyse se base généralement sur un état de l'art classique des technologies ou des brevets existants (Lee, Yoon, & Park, 2009), ou à partir de méthodes de créativité et de prospective plus structurées, telle que la méthode DELPHI (Linstone & Turoff, 1975) ou TRIZ (Altshuller, 1999).
- **Dimension technique et d'usage instable** : Dans le dernier cas, celui d'une double instabilité, il est possible d'opter pour des méthodes permettant l'exploration simultanée des dimensions techniques et d'usages, telle que la méthode D4 (Gillier & Piat, 2008; Gillier & Piat, 2011; Piat, 2005) qui se base sur la théorie CK de l'école des mines (Hatchuel & Weil, 2003). Elle repose sur 4 grandes étapes qui fondent son acronyme : D1 – « *Déconstruction* » (analyse des propriétés élémentaire de la technologie), D2 – « *Déclinaison* » (regroupement des propriétés en fonctions génériques), D3 – « *Destination* » (translation des fonctions potentielles sur différents marchés), et D4 – « *Décision* » (réévaluation des caractéristiques de la technologie).

Une fois les phases amont du projet de conception passé, de nombreux auteurs ont souligné le rôle important du PROTOTYPAGE RAPIDE dans le contexte des technologies émergentes et des applications innovantes (Anastassova, Mégard, & Burkhardt, 2007; Thomke, 1998). En effet, comme nous l'avons précédemment souligné, le principal défi dans ce cadre est de cerner le plus précisément possible les besoins des utilisateurs. Cette tâche doit être entièrement circonscrite avant de décider de déployer entièrement le développement des technologies associées, souvent très coûteuses en temps et moyens. En effet, à cause de leurs caractères innovants, il est rare que les technologies émergentes répondent à des besoins « *conscients* » (Robertson, 2001). Dans la majorité des cas, il s'agit plutôt de réalisations techniques en recherche d'application, qui répondent donc à des attentes « *latentes* » (Sperandio, 2001) car souvent ignorées ou mal traitées par les technologies précédentes. Cela rend difficile l'analyse de l'activité dans le sens où elle se construit au fur et à mesure que le projet d'innovation se définit (Brangier & Bastien, 2006). L'utilisation du PROTOTYPE SEMI-FONCTIONNEL peut répondre à ce problème pour plusieurs raisons. (i) Sa mise en place est relativement peu coûteuse et les fonctions les plus avancées peuvent être simulées à partir de la méthode du MAGICIEN D'OZ (Molin, 2004). (ii) Deuxièmement, il permet le recueil de données objectives qui, couplées à des données subjectives, offre la possibilité d'accéder à des données fines sur l'expérience d'interaction. (iii) Troisièmement, dans le contexte de l'innovation, il est plus facile pour l'utilisateur de formuler des besoins quand la technologie est arrivée à un niveau de développement suffisant. Il est en effet difficile pour des utilisateurs d'exprimer des besoins au sujet d'un outil et de situations d'usage qu'ils n'arrivent que partiellement à imaginer (Karsenty, 2004). Là où les méthodes classiques en IHM permettent aux utilisateurs l'élicitation de besoins

Figure 45 – Ecran de visualisation de l'audience distante avec simulation par Wizard of Oz des participants

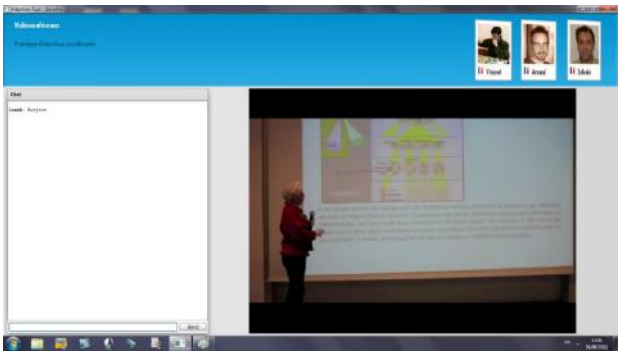


Figure 46 – Interface de vidéoconférence avec simulation par Wizard of Oz des participants et du présentateur

conscients sans nécessiter d'artefacts avancés, dans le cadre des applications innovantes −dont on fait l'hypothèse a priori qu'elles répondent à des besoins (Sperandio, 2001), donc ici à des « *besoins inimaginés* » (« *undreamed of requirements* » ; Robertson, 2001)− le prototype permet de faciliter la visualisation des applications potentielles de la technologie innovante instanciée (Anastassova, et al., 2007). (iv) Enfin, cette méthode permet de simuler l'interaction de l'utilisateur avec d'autres participants, ce qui est très important dans notre contexte, c'est-à-dire celui des applications de communication innovantes. La simulation de différents systèmes innovants, par l'utilisation de PROTOTYPAGE INTERACTIF et de la technique du MAGICIEN D'OZ (Figure 45 et 46), s'est particulièrement illustrée ces dernières années au sein des Bells Labs (Gonguet, Martinot, Rodio, & Hiribarren, 2013). Elle a permis, entre autres, de soulever différents problèmes posés par la capture et la représentation des émotions pour les systèmes de communications immersifs (Gonguet & Rodio, 2011) ; ou encore de mesurer l'impact du type d'interface et du style de présentation sur l'expérience utilisateur des auditeurs distants dans le contexte de la vidéo-conférence (Rodio & Gonguet, 2011).

Il est important de souligner que, dans le cadre d'applications innovantes, l'analyse et la formalisation des besoins se poursuivent souvent très tard dans la conception. En effet, l'analyse de la demande en début de projet –considérée souvent comme le pilier de l'intervention ergonomique− perd beaucoup son intérêt dans des environnements très disruptifs, même après plusieurs reformulations, car les besoins se « construisent » souvent au fil de la maturation conceptuelle même du projet. Ainsi, d'autres approches, plus constructivistes, ont été mises au point afin de prendre en compte cette émergence des besoins disruptifs lors de la confrontation des utilisateurs à des artéfacts évoluant au fil de la conception (Beguin, 2007; Falzon, 2005). D'un point de vue plus quantitatif, l'avancée dans le projet bénéficie également aux tâches d'évaluation car le degré de maturation de l'artéfact, en matière de fidélité visuelle et d'interaction, contribue à améliorer la validité de la mesure (Brajnik & Giachin, 2014).

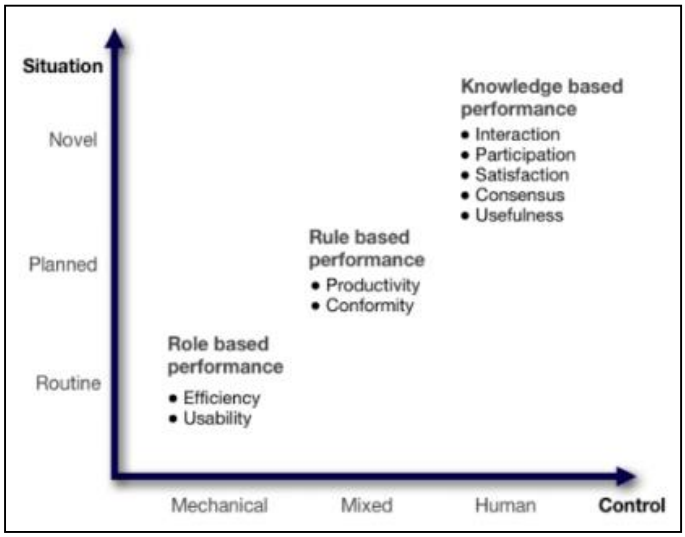


Figure 47 – Type d'évaluation et de contrôle en fonction du degré de nouveauté de la situation (adapté de *Antunes et al. 2012*)

Enfin, il est important de souligner que dans le cadre d'activités novatrices, les métriques utilisées pour l'évaluation se centrent plus généralement sur l'expérience humaine plutôt que sur les caractéristiques techniques des systèmes. Le modèle de Reason (2008), basé sur une extension des travaux de Rasmussen et Jensen (1974), permet d'en expliquer les raisons. Ce modèle a d'ailleurs été repris par Antunes et al. (2012) pour situer les méthodes d'évaluation des systèmes collaboratifs par rapport au degré de nouveauté des situations d'interaction. En résumé, ce modèle stipule que les modes de contrôle (pouvant être « *Mécanique* », « *Mixte* » ou « *Humain* ») varient en fonction du degré de nouveauté de l'activité (pouvant être « *Routinière* », « *Planifiée* », ou « *Nouvelle* »). Antunes et al. (2012) analyse ensuite le lien entre ces deux facteurs et le type d'évaluation associé (Figure 47) :

- **Évaluation d'une situation « Routinière » :** dans ces circonstances particulières, l'activité est bien cernée par les opérateurs et les détails techniques du système sont très bien formalisés. Le contrôle de l'activité est d'ordre mécanique (nombre d'actions, de mouvements de souris, de pressions clavier, …). L'évaluation se base donc généralement sur des métriques de performance, d'utilisabilité, de fiabilité, … Elle s'effectue à partir d'un contrôle des tâches usuelles, réalisé selon une séquence d'actions prédéfinies, et imposée par la technologie. Antunes et al. (2012) donnent comme exemple de méthodes, utilisées dans ce scénario, les MODELES DE PERFORMANCE HUMAIN, tels que le *« Keystroke-Level Model »* (Card et al., 1980b) ou le « *Human-Performance Models* » (Antunes et al., 2006). On peut y incorporer également les TESTS D'UTILISABILITE QUANTITATIFS dont l'objectif est d'affiner les performances d'un système déjà opérationnel (selon une approche semblable à de l'ergonomie de correction).
- **Évaluation d'une situation « Planifiée » :** dans ces circonstances, l'activité présente des familiarités avec des situations préexistantes. Il s'agit donc d'une activité, ni tout à fait nouvelle, ni tout à fait similaire à d'autres existantes. De ce fait, des règles de résolution ou des normes ont déjà été mises au point, durant le développement d'une expertise. L'évaluation se base donc généralement sur des règles ou des procédures visant la mise en conformité de l'activité conçue selon des plans préétablis. Ces méthodes donnent aux acteurs une plus grande latitude décisionnelle malgré une intervention encore contrainte par un environnement technique imposé. Antunes et al. (2012) donnent dans ce scénario comme exemple de méthodes les EVALUATIONS HEURISTIQUES ou les méthodes de WALKTHROUGH (Baker et al., 2002; Pinelle & Gutwin, 2002). On peut y incorporer également les TESTS D'UTILISABILITE QUALITATIFS dont l'objectif est de cerner les exigences et les désirs d'utilisateurs confrontés à un artéfact en cours de formalisation (selon une approche semblable à de l'ergonomie de conception)

- **Evaluation d'une situation nouvelle:** dans le dernier cas, l'activité en question apparait comme non-familière et repose sur des choix techniques encore peu circonscrits. L'immaturité des objets techniques, en cours de déploiement, biaise toute approche de contrôle purement mécanique et le caractère innovant de la situation ne permet pas d'y appliquer un savoir expert ou des normes techniques précédemment développées. Le contrôle y est donc principalement humain. Sans repères, ni règles de conception, l'évaluation ne peut que se baser sur des indicateurs de qualité rendant compte de l'expérience utilisateur, telles que la satisfaction ou l'utilité perçue, obtenue à partir d'une exploration ouverte d'un objet en cours de formalisation. Antunes et al. (2012) donnent comme exemple de méthodes le TECHNOLOGY TRANSITION MODEL (Briggs et al., 1998) et la méthode de PERCEIVED VALUE (Antunes et al., 2006), qui reposent tous deux sur l'appréciation subjective par les utilisateurs des facteurs affectant la valeur et la pertinence perçue d'un système. On peut y incorporer également toutes les méthodes visant à exprimer ou mesurer l'expérience utilisateur, de manière générique ou selon des dimensions plus spécialisées.

En résumé, le caractère innovant de certaines situations d'interactions nous oblige à passer de stratégies d'évaluations centrées sur l'instrument à des stratégies centrées sur l'humain. Ainsi, pour Antunes et al. (2012), il y a dans le domaine de l'évaluation des systèmes collaboratifs modernes, un glissement du paradigme « *instrumental* » vers le paradigme « *inter-subjectiviste* ». Cela favorise ainsi l'utilisation de méthodes centrées sur l'expérience utilisateur car conçue à la base pour tenir compte de la qualité subjectivement perçue par les utilisateurs du système étudié. De plus, les méthodes UX, par leur ouverture et leur flexibilité, offrent des avantages non négligeables dans le domaine de l'innovation où il convient d'intervenir tôt dans la conception et rapidement. Cela amena donc à favoriser, dans la conception d'applications numériques innovantes, l'utilisation de méthodes plus qualitatives que quantitatives.

Le Web social et le retour de l'analyse quantitative

Le chapitre précédent nous a montré la tendance actuelle des innovateurs à se concentrer sur les phases amont de la conception, en usant de technique plus qualitative et ouverte. Néanmoins, il existe une révolution méthodologique encore plus grande, en aval cette fois. Amorcé timidement dès la naissance du web social, ce mouvement a commencé à montrer le bout de son nez il y a peu, et s'amplifiera largement dans les années à venir. En effet, Le Web 2.0, et surtout les technologies qui vinrent ensuite, est en train de transformer les utilisateurs en acteurs à part entière du processus d'innovation. Cela va bouleverser en profondeur la manière dont nous évaluerons les applications.

Les données utilisateurs et l'avènement de l'économie de la gratuité

Peu après l'éclatement de la bulle Internet, en 2001, nous avons assisté rapidement à mutation globale de ses fondations. À partir d'angles d'analyse différents, des études ont permis de prendre conscience de la naissance du web social. Des textes fondateurs comme le « *Cluetrain*

Manifesto » (Levine, 2000), des conférences comme celles de Tim O'reilly [38] et les succès d'applications telles que Napster, Myspace ou Facebook, ont remis en cause la vision réductrice d'Internet comme simple média de distribution. Le Web 2.0 a ouvert une nouvelle dimension d'Internet, celle de la participation des utilisateurs ordinaires. Il est intéressant de noter que ce mouvement est, dans un certain sens, à l'opposé de celui de la société des masses et des *mass medias,* qui ont dominé nos sociétés durant plus de 150 ans. En effet, la production de contenus avait toujours été possédée par de grandes sociétés disposant de capitaux et de moyens de distribution colossaux. Avec Internet, il est devenu possible de produire et diffuser du contenu personnel avec une facilité inédite jusqu'alors. Blank & Reisdorf (2012), en reprenant un principe de l'économie de la permanence de Ghandi, parle du passage de *« la production de masse à la production par les masses* » (p. 541). Ce changement a, dans un premier temps, fortement déstabilisé les oligopoles traditionnels. On se souvient en effet de l'affrontement des labels musicaux avec Napster au début des années 2000. On sait maintenant qu'aucun des géants numériques d'aujourd'hui n'aurait pu naître s'il ne s'était pas appuyé −au contraire des anciens oligopoles− sur les données engendrées par les utilisateurs. Ce véritable *« or numérique* », capté et centralisé par différentes plateformes, a fait la force d'entités telles que Facebook et Youtube. On peut dire même que ces géants sont nés grâce à l'utilisation de ces « *User data* », éclipsant progressivement tous les autres qui s'acharnaient encore à utiliser les anciens modèles de création et de distribution de bien numérique. De plus, leur position de précurseur leur a permis d'amasser une quantité colossale de contenus, les rendant ainsi de plus en plus attractifs grâce à un effet de gravité. Nous avons vu dans les chapitres précédents en quoi l'exploitation de l'effet de réseau représente un avantage compétitif dans l'économie numérique. Cet effet d'autant plus accentué par les systèmes conçus pour capter, organiser et valoriser les données issues du suivi régulier et systématique de l'activité des utilisateurs. De ce fait, Blank & Reisdorf (2012) définie le Web 2.0 comme « *l'utilisation d'internet pour fournir des plateformes à travers lesquelles les effets de réseau peuvent émerger* » (p. 539). Mais la quête des « *User Data* » est guidée également par l'existence nombreux autres avantages concurrentiels, inhérents aux propriétés du web moderne, ce qui en fait LA priorité des acteurs actuels du digital.

Ainsi, il est important de comprendre que les données utilisateurs sont devenues le flux essentiel qui irrigue toute l'économie numérique. Elles sont d'origines très variées : historiques de recherches, de consultations et d'achats; contacts et données partagées sur les réseaux sociaux ; localisations géographiques, paramètres médicaux et physiologiques, etc. Elles peuvent être recueillies à partir de données soumises explicitement par l'utilisateur, de traces d'utilisation (clic, navigation, défilement de pages, …) ou de données inférées à partir de traitements ou de recoupements. Elles permettent, entre autres, de mesurer et d'améliorer les performances d'un service, d'en personnaliser le contenu, de prendre des décisions stratégiques (pouvant donner naissance à d'autres solutions) ou d'être valorisées auprès de tiers. En résumé, les données utilisateurs sont les leviers qui permettent aux entreprises modernes du numérique d'atteindre de hauts niveaux de profitabilité à moindre coût. La collecte des données utilisateurs contribue ainsi à l'émergence d'un nouveau modèle, celui du travail gratuit ou de l'économie de la gratuité

[38]Tim O'reilly est l'initiateur de nombreuses conférences sur le Net depuis 2001 et l'inventeur du terme Web 2.0.

(Collin & Colin, 2013). En effet, les utilisateurs, attirés par la qualité des interfaces et l'effet de réseau, deviennent, au travers des traces laissées, des auxiliaires de la production et de la création de valeurs, sans pour autant toucher de contrepartie monétaire. Nous observons donc un dépassement de la théorie de la firme telle que formulée par Ronald Coase, en 1937, qui suggère qu'une entreprise doit choisir entre la sous-traitance et le recrutement de salariés. Ici, une troisième alternative apparaît : celle de la mise au point d'un service qui suggérerait à ses utilisateurs une activité dont les externalités positives viennent s'incorporer à la chaîne de production **gratuitement**. C'est l'absence de rémunération des utilisateurs pour leurs données qui explique actuellement les rendements d'échelle exponentiels, propres à l'économie numérique. Elle réduit les coûts de sous-traitance (« *outsourcing* »), préférant l'entraide mutuelle et gratuite des clients de leurs services (« unsourcing »[39]). Ainsi, la faiblesse des coûts d'exploitation, résultant de la délégation de certaines tâches aux utilisateurs, permet la gratuité de la plupart des services en ligne, permettant en retour de démultiplier l'attraction du service. Cette première phase de traction des utilisateurs et de leurs données est suivie ensuite d'une phase de valorisation financière de ce capital. Cette dernière peut prendre plusieurs formes. La première manière, capitalisant sur un processus de fidélisation, consiste à faire payer l'utilisateur une fois que ce dernier s'est habitué à utiliser le service (Anderson, 2009). C'est le cas par exemple de l'ancien modèle de diffusion de logiciel « *shareware* »[40], et, plus récemment, du modèle commercial « *freemium* »[41]. En effet, l'amélioration des transactions financières numériques −via des interfaces plus ergonomiques et sécurisés− et la possibilité de segmenter et monétiser finement l'offre de service, −via l'utilisation du micropaiement et des monnaies virtuelles− permettent de favoriser ces stratégies commerciales, aux dépens d'offres plus classiques. La deuxième manière de valoriser monétairement ce capital est de procéder au financement du service par une autre face du modèle d'affaire. La méthode la plus courante est d'avoir recours à la publicité, dont les traces utilisateurs permettent, de plus, un ciblage d'autant plus précis. D'autres stratégies de valorisation ont également été envisagées, comme le fait de mettre ces données à la disposition de développeurs pour l'amélioration de services encore immatures (ex : les anciennes applications Google Labs), de développer de nouveaux produits (comme 23andMe qui va utiliser sa base de donnée génétique pour créer de nouveaux médicaments), de permettre de réaliser des gains d'achat ailleurs (ex : le moteur de recommandation d'Amazon) ou tout simplement de les vendre ou de les louer à des tiers (comme le fait TripAdvisor). Ainsi, si le produit que vous utilisez est gratuit, il y a fort à parier que ce soit **vous** le produit[42]. Toutes ces stratégies « *Data to Value* » ont été fortement utilisées par le GAFA (Google, Apple, Facebook et Amazon) et ont grandement contribué à leur domination du marché numérique. Ces derniers totalisent en 2014 316 milliards de dollars de chiffre d'affaire et emploient ensemble 252 000 personnes, soit l'équivalent du PIB du Danemark, avec 10 fois moins de personnes (Fabernovel, 2014).

---

[39] Terme utilisé sur le blog de « The economist », dans un article de 2012 intitulé « Outsourcing is so last year »

[40] Le « *Shareware* » revoie généralement à une application pouvant être utilisé gratuitement durant une certaine période de temps. Après ce délai, l'utilisateur doit rémunérer le/les auteur(s) s'il veut continuer à utiliser le logiciel.

[41] Le freemium, contraction de « free » et « premium » est une stratégie commerciale associant une offre gratuite, en libre accès, et une offre « Premium », haut de gamme, en accès payant.

[42] Selon le slogan célèbre de l'agence de communication audiovisuelle Adesias

La deuxième dimension de la réflexion autour des modèles 2.0, à côté de ses grands avantages économiques, est celle de la transformation même de la création de connaissances, dont Wikipédia a été pour beaucoup l'exemple de réussite. En effet, en mobilisant une foule énorme de personnes ou de données, il est possible d'arriver à des résultats impossibles à réaliser par des moyens conventionnels. Nous avons déjà cerné cette dimension sous la notion de « *sagesse des foules* » (Surowiecki, 2005). Sa mise en évidence ne remonte pas d'hier et le célèbre statisticien Galton lui-même en avait déjà fait part dans un article de Nature datant de 1907, sous la forme d'une anecdote vécue lors d'un concours d'une foire paysanne à Plymouth. L'objectif du concours était de deviner le poids en viande d'un bœuf présenté vivant. Galton constata que la moyenne des réponses des 787 participants était de 1197 livres, soit une estimation à seulement une livre près de la réponse exacte : 1198 livres ! (Galton, 1907). L'exploitation industrielle de cette propriété, issue de la collaboration de masse, est désignée de nos jours par le néologisme « *crowdsourcing* », s'inspirant du terme « *outsourcing* » (sous-traitance) et de « c*rown* » (la foule). Nous devrons ce terme à Jeff Howe qui explique que les sauts technologiques et la diffusion des outils informatiques bon marché ont fortement réduit certains écarts entre professionnels et amateurs, ce qui permet aux entreprises de profiter des talents de la population, notamment par ce crowdsourcing (Howe, 2006). Plus précisément, nous pouvons définir ce nouveau mouvement comme « *un modèle de production et de résolution de problèmes en ligne et massivement distribué* » (Brabham, 2008). Ainsi, pour certains, la foule sera l'acteur au centre des *business models* de demain (Dawson & Bynghall, 2012). Dans cet ordre d'idées, le site internet crowdsourcing.org[43] classe ces nouvelles initiatives commerciales dans sept grandes catégories industrielles : le financement par la foule (« *Crowdfounding* »), le travail dans le Cloud (« *Cloud Labor* »), la créativité collective (« *Collective Creativity* »), l'innovation ouverte (« *Open Innovation* »), l'intelligence collective (« *Collective Knowledge* »), le « *Community Building* » et l'engagement civique.

L'optimisation de l'expérience utilisateur comme levier pour engager les utilisateurs

Pour se faire une place dans l'écosystème du Web 2.0, il ne suffit pas seulement de mettre en ligne une plate-forme, car il faut encore que les utilisateurs l'adoptent en nombre suffisant pour que l'effet de réseaux se produise. En effet, Internet est jonché des cadavres de plates-formes ayant eu la prétention d'être la prochaine révolution numérique. Ainsi, si l'adoption des utilisateurs est un facteur clef du Web 2.0 (Blank & Reisdorf, 2012), elle est difficile à anticiper. Deux solutions sont souvent employées par les grands groupes pour maximiser les chances que les utilisateurs utilisent leurs services. La première −la plus sûre mais également la plus coûteuse− consiste à acheter les start-up prometteuses pour compléter leur portefeuille d'offres, comme l'ont fait dernièrement Facebook et Microsoft pour Occulus VR[44] et Minecraft[45], pour plusieurs milliards de dollars chacun. La deuxième solution est l'amélioration continue des services en investissant lourdement dans l'expérience utilisateur. En effet, là où le modèle

[43] Voir l'infographie du site crowdsourcing.org « Crowdsourcing Industry Landscape » publiée en Mai 2011

[44] https://www.oculus.com/

[45] https://minecraft.net/

traditionnel se base fortement sur la publicité (par une diffusion pyramidale, principalement télévisée) pour créer une image de marque positive, le modèle du web 2.0 se base sur la qualité de l'expérience utilisateur pour se propager grâce au bouche-à-oreille (par une diffusion en réseau, principalement par Internet). On constate par exemple que Google et Facebook investissent en moyenne 30% de plus que leurs concurrents dans la recherche UX et Amazon y consacre pas moins de 80 % de son flux de trésorerie opérationnel (Fabernovel, 2014). De même, Apple n'a dépensé qu'un demi-milliard de dollars en publicité en 2013 aux US alors que P&G (multinationale américaine spécialisée dans les biens de consommation courante et possesseurs de marque telles que Always, Gillette ou Oral-B) en a dépensé plus de trois (Fabernovel, 2014).

L'expérience utilisateur des applications en ligne a déjà grandement bénéficié de l'utilisation de nombreux leviers techniques et humains. Parmi les leviers techniques, les temps de réponse ont été grandement améliorés grâce aux technologies Adsl et 3G, la programmation asynchrone (ex : Node.js) et de nouvelles technologies de gestion de cache (ex : Redis). Un autre enjeu technique a été celui de la mise en page dynamique, indispensable dans le contexte actuel de diversification des terminaux. Cette approche, dite du « *Responsible Design* » (Knight, 2011; Marcotte, 2011), a été grandement facilitée par l'introduction d'HTML 5. Comme autre levier d'amélioration de l'expérience utilisateur, le courant du *Design* a été adopté, via l'amélioration de l'esthétique et de l'évocation émotionnelle, pour tisser des liens affectifs entre l'utilisateur et le produit. La facilitation de la contribution des utilisateurs par une bonne utilisabilité du service d'une part, et par la mise en place d'une participation récompensée d'autre part, a également été utilisée pour développer et maintenir l'attractivité des services (Nielsen, 2006). Cela peut être par l'obtention de traitements préférentiels, de badges ou de points (en ludifiant[46] la participation telle que sur le célèbre forum informatique *Stack Overflow*) ou par des récompenses sociales, telles que la célébrité, la notoriété ou la réputation (telle que l'on peut la voir sur YouTube, Quora ou Twitter). Des hébergeurs de contenus, comme Youtube, vont même jusqu'à monétiser la participation des contributeurs en fonction de l'audience des productions et des revenus publicitaires. Le concept de « *Gamification* » (Marache-francisco & Brangier, 2013) et de technologies « *Persuasives* » (Fogg, 2003; Némery, Brangier, & Kopp, 2011) ont ici toute leur place pour capter, retenir et motiver les utilisateurs à utiliser les services proposés. À côté de l'utilisation de ces boîtes à outils conceptuelles et techniques pour améliorer l'expérience utilisateur, les développeurs se sont approprié des méthodes de conception et d'évaluation spécifiquement adaptées à la philosophie du Web 2.0.

Les méthodes de conception et d'évaluation des systèmes numériques sous le prisme du Web 2.0

Nous avons vu que dans Web social (ou web 2.0), le consommateur n'est plus seulement le point d'arrivée de la production, mais un acteur à part entière dans le processus d'innovation. Le web permet la structuration de communautés permettant à des amateurs, portés par leurs passions, d'obtenir la reconnaissance pour leurs connaissances. Ce mouvement d'innovations

[46] La ludification (ou en anglais « gamification ») est le transfert des mécanismes du jeu dans d'autres domaines, comme les sites web ou des situations d'apprentissage ou de travail

Figure 48 – Exemple de structure conçue par les joueurs d'EverQuest Next Landmark

ouvertes prolonge le mouvement des FabLab[47], né dans les années 90 au MIT, et du « *Do-it Yourself* »[48] dans le domaine particulier du numérique. Ainsi apparaissent dans les entreprises de nombreux profils de « C*ommunity Manager* », ayant pour objectif de mettre en place des plates-formes participatives ouvertes en interne (employés) et en externe (consommateurs, usagers, grand public) pour favoriser l'innovation et de permettre la co-création de services, produits ou contenus. L'exploitation de la sagesse des foules est également à l'étude afin de décupler la créativité en amont des projets via l'utilisation de plates-formes en ligne (Yu, Kittur, & Kraut, 2014a, 2014b; Zhao et al., 2014). Ainsi, le mode de développement à ciel ouvert que l'on observe sur le web et dans le logiciel aujourd'hui change totalement les modalités de prise en compte des utilisateurs par les innovateurs, puisqu'ils sont intégrés dans la boucle très tôt (Boullier, 2010b), et selon des moyens techniques qui mettent directement en contact les développeurs web avec leurs usagers. Le Crowdfunding (ex : kickstarter) est un bon exemple d'outils permettant à des projets en phase amont de recevoir une validation des futurs utilisateurs et d'être financés en même temps. La motivation des utilisateurs à participer aux projets de développement est également de plus en plus exploitée par les concepteurs. En effet, dans le nouveau jeu vidéo EverQuest Next Landmark[49], un éditeur de contenu et des outils sociaux sont fournis aux joueurs afin de leur permettre de créer du contenu de jeu (paysage, villes, édifices, …) de leurs propres mains (Figure 48). Certaines de ces structures seront sélectionnées par Sony et incluses dans la version finale du jeu. Les joueurs participent ainsi à la conception sans autre récompense que de voir leur création figurer dans la version définitive du jeu. Un engouement similaire est identifiable pour participer aux versions de test des jeux à

[47] Les FabLabs sont des laboratoires collaboratifs où chacun est invité à se joindre à une équipe pour travailler, apprendre et mettre en œuvre ses idées pour arriver à l'élaboration d'un produit concret.

[48] Le « *Do-it Yourself* » (ou DIY) est un mouvement consistant à construire, modifier ou réparer des objets sans l'aide d'experts ou de professionnels

[49] https://www.landmarkthegame.com

sortir, que ce soit en beta fermé[50] ou même en pré-alpha, à acquérir au prix du neuf[51]! La startup Dandy [52]va encore plus loin, en proposant une plateforme de développement mobile « *crown-sourcée* » (crowd-sourcing), c'est-à-dire développé par les membres de la communauté même, et rétribué financièrement en fonction de leur niveau de contribution.

On constate ainsi dans le développement numérique une tendance lourde consistant à mettre le produit entre les mains des utilisateurs le plus tôt possible afin de profiter de sa participation, de son feedback et de son financement, via divers dispositifs en ligne. Suivant l'idéologie web 2.0 de la Beta perpétuelle (Bellucci et al., 2010), ces produits sont ensuite suivis et révisés continuellement tout au long de leur cycle de vie. Des mécanismes de feedback utilisateurs sont ainsi intégrés à l'intérieur même des programmes (pour les plus connus : *UserVoice*, *Usabilla* et *Get Satisfaction*). Ils permettent aux utilisateurs, sur un mode proche des évaluations formatives, de remonter à tout moment leurs appréciations, difficultés ou de suggérer de nouvelles fonctions.

À côté de l'internalisation, au sein des applications web, du feedback utilisateur **qualitatif et volontaire**, une autre tendance lourde se développe actuellement : dénommé par le passé « *Datamining* », puis depuis peu « *Big Data* », ce mouvement tente de valoriser **quantitativement** les traces d'interactions laissées **involontairement** par les utilisateurs. Le traitement de telles masses de données n'a été rendu possible que depuis peu, grâce aux progrès technologiques et techniques : diminution du coût de stockage des informations, mise au point de structures de données adaptées (telles qu'avec les bases de données non relationnelles) et d'algorithmes permettant un traitement massif de ces données. Le *Big Data* est un domaine d'activité en pleine explosion, d'autant plus que le nombre d'objets connectés et de capteurs disponibles ne va pas cesser d'augmenter. On estime ainsi que le potentiel d'exploitation de ces données personnelles à des fins commerciales pourrait s'élever à plus de mille milliards d'euros d'ici 2020, pour valoir 8% du PIB pour les pays du G20[53]. C'est dans ce contexte que le commissariat général à l'investissement a lancé un appel à projets *Big Data*, doté de 25 millions d'euros dans le cadre du programme des investissements d'avenir[54]. Il existe une règle très simple pour reconnaître un produit *Big Data :* c'est la règle des 3V (Manyika et al., 2011). Les trois V correspondent à :

- **Volume** : la masse de données est conséquente, de l'ordre du péta (10^{15}) ou de l'exaoctet (10^{18})
- **Variété** : les données sont de natures très diverses, (vidéos, logs, mails, images, …)
- **Vélocité** : le traitement des données est accéléré, voire parfois en temps réel

---

[50] La bêta fermée ou bêta privée est une version du jeu dans laquelle les personnes intéressées doivent s'inscrire au préalable ou sont contactées par les concepteurs du produit testé qui sélectionnent les candidatures.

[51] Jouant sur l'exclusivité de pouvoir découvrir un jeu avant sa sortie, la tendance du pré-alpha propose d'acheter l'accès à des jeux encore dépourvus de certaines fonctionnalités et contenant encore un nombre de bugs encore important, mais dont le financement participatif permet le prolongement du développement.

[52] https://angel.co/dandy

[53] Etude du Boston Consulting Group et de Liberty Global (Novembre 2012), nommé « *The Value of our Digital Identity* » : http://www.libertyglobal.com/PDF/public-policy/The-Value-of-Our-Digital-Identity.pdf

[54] Commissariat général à l'investissement (22 mars 2012). « Lancement de l'appel à projets consacré au Big Data ». http://investissement-_avenir.gouvernement.fr/

Le Big Data apporte au marketing une puissance et une précision sans précédent : personnalisation en ligne, publicité en temps réel, analyse de sentiments, etc. Les méthodes abondent pour identifier au plus près les besoins de l'utilisateur et lui communiquer ainsi, en temps réel, une réponse à ses attentes.

Le Big Data réhabilite aussi habilement l'évaluation quantitative dans le cycle de conception. L'utilisation de ces données permet de nombreuses applications, allant de l'amélioration continue à la personnalisation de services. En effet, le suivi régulier et systématique de l'activité des utilisateurs peut servir à mesurer et améliorer les performances d'une application par des corrections et améliorations ciblées et bien choisies. Ce pilotage de la performance d'un service, par des indicateurs clefs issus de la collecte de données utilisateur, est appelé « *Growth Hacking* », dont la compétence est située par ses afficionados à la croisée du marketing du développement. La philosophie des « *Growth Hacker* » est qu'un produit doit être outillé de telle sorte qu'il soit à tout moment testable et évaluable, afin d'être constamment amélioré, jusqu'à atteindre une adoption massive (Holiday, 2014). La finesse de l'analyse peut même aller jusqu'à l'observation en temps réel de nombreux éléments individuels du service (articles, fonctions, pages, …). Par exemple, de nombreux sites d'informations −incluant Gawker, Forbes et le New York Times− utilisent d'ores et déjà des systèmes d'analyses en temps réel de l'attention portée sur leurs articles, via des outils comme Chartbeat (Figure 49).

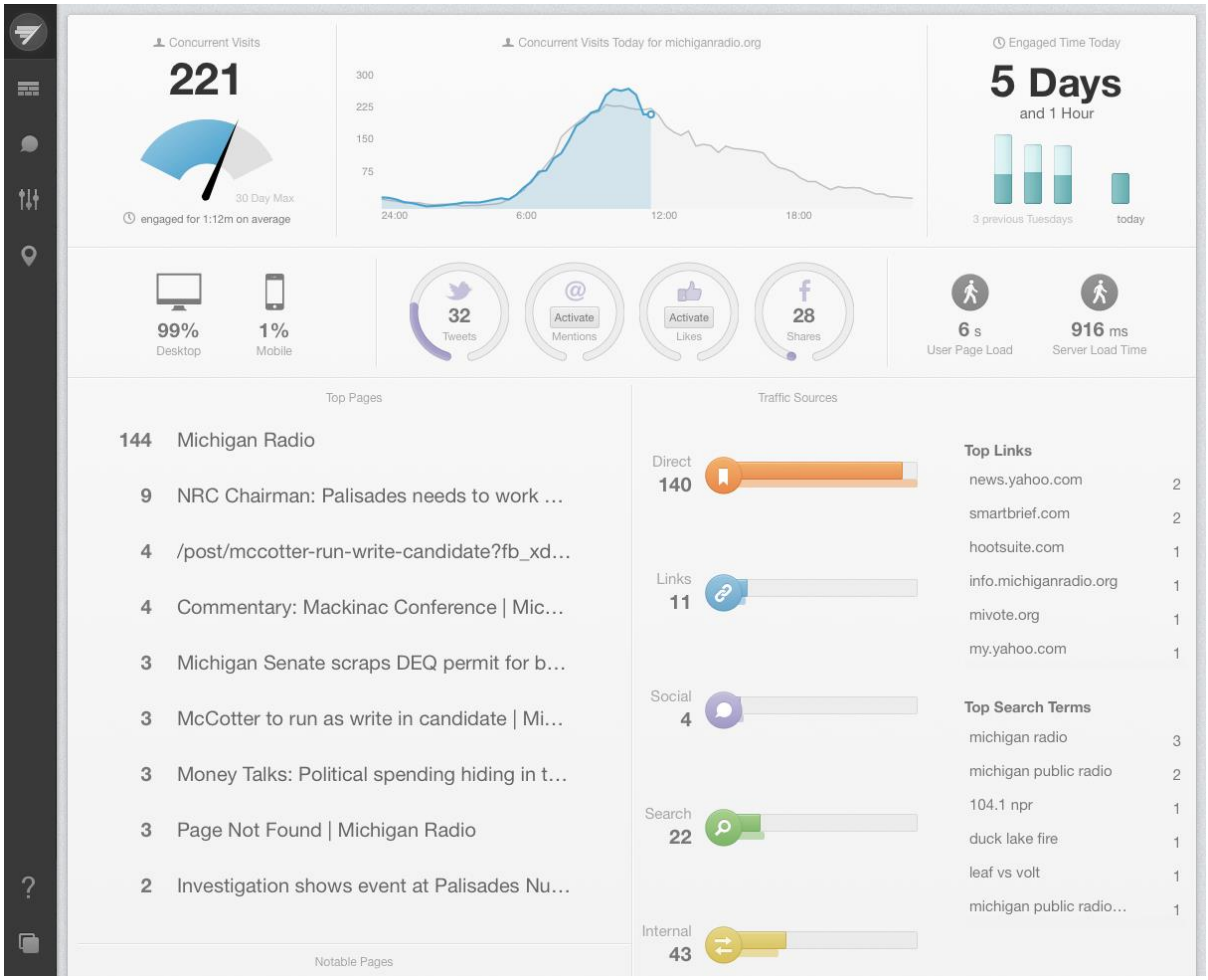


Figure 49 − Exemple de tableaux de bord fourni par Charbeat pour mesurer en temps réel l'attention des utilisateurs pour un site internet donné

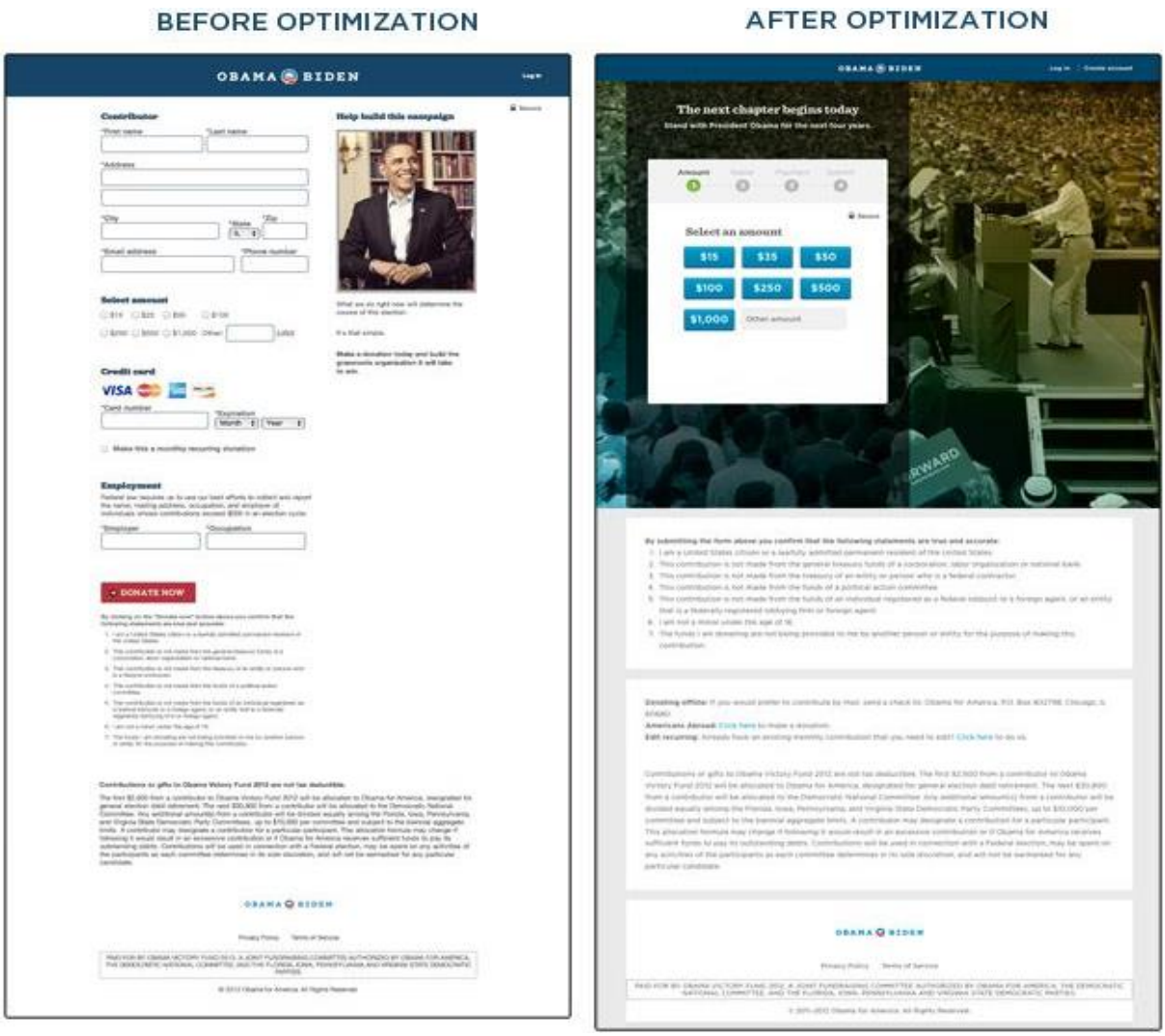


Figure 50 − Le site de la campagne Obama avant et après son optimisation par A/B testing (tiré du site kylerush.net)

Ces dispositifs d'évaluation interne permettent également de réhabiliter une méthode souvent écartée par le passé, car très coûteuse en temps et en moyens : l'A/B TESTING. Le web pouvant décliner à faible coût différentes versions d'un même produit et profiter largement d'un panel d'utilisateur gigantesque, il redevient possible de trancher facilement entre plusieurs solutions concurrentes, en utilisant Internet comme plate-forme de test. Les données issues de l'activité des utilisateurs peuvent ainsi faire l'objet d'une analyse statistique pour identifier l'interface qui maximise le plus l'acquisition de nouveaux visiteurs, leur niveau d'engagement, le montant du panier moyen dans les applications de vente, etc. L'A/B TESTING à grande échelle est d'ailleurs régulièrement utilisé par Facebook, faisant coexister en permanence plusieurs versions de son application, dont elle mesure les performances relatives pour inspirer l'évolution future de son design (Jana, 2011). D'une manière similaire, 240 A/B tests successifs ont permis d'augmenter de 49% le taux de conversion des visiteurs sur la page d'appel aux dons de la campagne présidentielle de Barack Obama (Figure 50) en 2012[55].

L'automatisation de l'évaluation quantitative dans le domaine du *Big Data* a permis également de résoudre un problème qui a émergé avec le Web 2.0 : l'évaluation des contributions

[55] Selon l'article de Kyle RUSH, « Meet the Obama campaign's $250 million fundraising platform »

utilisateurs. En effet, le glissement du Web 1.0 au Web 2.0 fit passer d'une logique de la page Web (issues d'administrateurs/rédacteurs) à une logique de contenus utilisateurs (Wang & Chiu, 2011). L'évaluation de ces contenus utilisateurs est importante car ces nouveaux services (souvent gratuits) vivent du haut niveau de leur fréquentation, ce qui les astreint à veiller à la qualité des données. Or, ces contributions sont si volumineuses qu'elles obligent à utiliser des évaluations automatisées, via des algorithmes spécialisés. Cette automatisation est vitale, d'autant plus que la mesure de l'audience dans le mouvement de digitalisation des médias et de l'explosion de la participation constitue un enjeu de plus en plus important (Jennes & Pierson, 2011).

Enfin, la grande quantité d'informations disponibles sur un utilisateur peut servir à personnaliser le service à son attention, en fonction de sa personnalité, de ses capacités et de ses besoins. Ces informations permettent au système de lui faire de bonnes recommandations, de lui épargner certaines étapes du parcours, ou, au contraire, d'approfondir un sujet en fonction de son expertise. Cette étape nous rapproche de l'application intelligente, selon l'approche symbiotique de Brangier et al. (2009), dans le sens où le système, étant capable d'une évaluation du statut cognitif, émotionnel et comportemental de l'utilisateur en temps réel, peut adapter l'interface pour améliorer de manière significative l'exécution de la tâche avec l'utilisateur.

En conclusion, nous voyons que les développeurs du Web 2.0 ont intégré dans leurs applications toute une série d'outils de recueil du feedback utilisateurs, à partir de traces involontaires laissées lors de l'interaction ou d'expressions plus volontaires, permettant une évaluation formative ou sommative de leurs solutions tout au long du cycle de vie de l'application. La réduction de la viscosité entre l'utilisateur et les technologies numériques, par toute une série de passerelles, permet une amélioration des services de manière continue et à faible coûts.

LA MESURE DE L'EXPERIENCE UTILISATEUR A L'ERE DU BIG DATA

La prise en compte du contexte général, qui entoure la conception et l'évaluation des applications de communication, de ses débuts jusqu'à nos jours, nous permet d'appréhender les défis méthodologiques qui se présentent aujourd'hui.

Nous avons vu que le domaine de l'évaluation des systèmes numériques s'est constamment adapté afin de répondre à de nouveaux enjeux, issue de toute une série de transformations techniques, sociétales et économiques du domaine. En effet, au début de l'informatique, dans les années 60, cette évaluation fut encore informelle et naïve. Puis, dans les années 70, quand l'informatique rejoint l'entreprise, afin d'optimiser ses processus de production, elle devint plus rigoureuse et centrée sur la performance. Dans les années 80, quand le milieu de la conception informatique fut plus concurrentiel et réactif, elle s'arma de méthodes plus souples, économiques et itératives. Elle se focalisera également de plus en plus sur la facilité d'utilisation, quand l'utilisateur final deviendra progressivement le principal acheteur. Enfin, depuis la fin des années 90, lors du débordement de l'informatique en dehors des murs de l'entreprise pour investir les loisirs, l'évaluation va se recentrer sur l'expérience utilisateur, entraînant un enrichissement des méthodologies et des disciplines académiques associées.

En ce moment même, deux autres grandes transformations du domaine modifient les méthodologies privilégiées pour la conception et l'évaluation des systèmes numériques. D'une part, l'accélération du renouvellement de l'offre, par la mise en place d'une économie de l'innovation, va concentrer les efforts en amont de la conception, en s'appuyant sur des méthodes qualitatives, alliant créativité et anticipation des besoins. D'autre part, la mise en ligne précoce des produits, couplée à des mécanismes internes de recueil du feedback utilisateurs (majoritairement quantitatif), va permettre de poursuivre à faibles coûts l'adaptation des services en fonction des exigences des utilisateurs ciblés. Pour résumer, on constate une prise en compte de l'utilisateur de plus en plus en amont, une réduction temporelle des phases de conception, avec une utilisation prépondérante des méthodes qualitatives ; puis, par une mise à disposition précoce du produit, accompagné d'une amélioration continue via des mécanismes de recueil internes, une explosion des évaluations quantitatives, routinières et automatisés de l'expérience des utilisateurs. Ainsi, il est important de se demander quelle place reste-t-il à l'ergonome dans la configuration actuelle.

CHAPITRE 6 : LA PLACE DE L'ERGONOMIE MODERNE DANS LE CYCLE DE CONCEPTION EN IHM

« *Mesurer exactement un objet fuyant ou indéterminé, mesurer exactement un objet fixe et bien déterminé avec un instrument grossier, voilà deux types d'occupations vaines que rejette de prime abord la discipline scientifique* »

Gaston Bachelard (1970).

L'accaparation du champ par les designers en amont et par les développeurs en aval

Nous constatons premièrement que le travail de « *conceptualisation* » en phase amont a été, en grande partie, accaparé par les designers. Bien qu'aucune définition universelle du design n'ait pu mettre d'accord l'ensemble des protagonistes de cette discipline, la plus citée est celle de l'International Council of Societies of Industrial Design (ICSID) : « *Le design est une activité créatrice dont le but est de déterminer les qualités formelles des objets produits industriellement. Par qualité formelle, on ne doit pas seulement entendre les qualités extérieures, mais surtout les relations structurelles et fonctionnelles qui font de l'objet une unité cohérente, tant du point de vue du producteur que du consommateur* » (C. Bouchard, 1997). Cette définition souligne le rôle du designer comme le garant de la **cohérence** du produit, ce qui résonne actuellement avec la visée **holistique** de l'étude de l'expérience utilisateur. La vision française du design rapproche également beaucoup la pratique du designer à celle de l'art appliqué, dont l'esthétique et l'hédonisme constituent leurs préoccupations majeures. L'esthétisation des objets et des expériences est particulièrement en phase avec les tendances de la société actuelle (Michaud, 2003). De plus, la production en masse d'artefacts artistiques hautement différenciés est devenue rentable, car les modèles de distribution moderne permettent la mise en place d'une économie « *de la longue traine*[56] ». On constate également que la vision de la conception « *artistique* », en opposition à la conception « *centrée utilisateur* », devient séduisante pour un grand nombre d'acteurs, car, partant d'une inspiration personnelle, elle est moins coûteuse en ressources humaines et temporelles. Enfin, d'autres praticiens du design voient pour objectif principal de leur discipline l'exploration des futurs possibles, en opposition directe avec d'autres orientations académiques, focalisées historiquement sur l'existant (Lowgren, 2013). Ainsi, le design, porté sur une approche qualitative, holistique et anticipative, apparaît pour beaucoup comme une approche fortement opérante pour intervenir en amont et durant les premières phases de conception, d'autant plus pour les projets focalisés sur l'innovation.

D'un autre côté, peu de temps après la conception du premier prototype fonctionnel, nous constatons que la pratique récente conduit à ce que le produit numérique soit déployé

[56] L'expression longue traîne (« *long tail* » en anglais) désigne la stratégie de vente consistant à proposer une grande diversité de produits, chacune représentant une faible demande, mais qui, collectivement, représente une part de marché égale ou supérieure à celle des best-sellers. Cet stratégie a été rendu accessible par les possibilités d'Internet et est utilisé massivement par des acteurs comme Amazon ou Netflix.

rapidement sur le web, afin de prendre en charge la suite de son développement, en fonction du retour des utilisateurs à son contact. En effet, les traces récoltées, laissées dans l'environnement Internet moderne, permettent (entre autres) une analyse riche de l'interaction et de l'expérience utilisateur, sans faire appel nécessairement à un ergonome. Mais ce bouleversement historique, consistant en une disponibilité massive et facilement accessible de données sur le comportement humain (Shneiderman, 2011), permet potentiellement de révolutionner bien d'autres pratiques. La quantité phénoménale de données disponibles aux chercheurs en sciences humaines est comparable à la révolution du microscope pour les biologistes (King, 2011). En effet, là où certains chercheurs basaient leurs travaux sur des enquêtes ne dépassant généralement pas le millier d'individus tirés au hasard, il est maintenant possible de collecter plus de cent millions de contributions sur les médias sociaux par jour, et d'en extraire l'information utile à partir de nouvelles méthodes d'analyse de texte automatique (Hopkins & King, 2010). Des questions portant sur le comportement individuel, coopératif ou social, sont ainsi traitées à l'intérieur de disciplines nouvelles, basé sur des méthodes propres, tel que les « *humanités numériques* » (Boullier & Lohard, 2012) du Médialab de Sciences Po ou de l' « *Human-Community Interaction* » de Ben Shneiderman (Shneiderman, 2011). De même, des champs dont la coloration méthodologique était très qualitative, tels que les CSCW, se tourne progressivement vers le quantitatif avec l'accessibilité à de gros volume de données comportementales sur internet (Grudin & Poltrock, 2012).

Néanmoins, la première communauté de chercheurs à s'être investie dans ce champ sont les promoteurs historiques de la « *science du Web* » (Hendler, Shadbolt, Hall, Berners-Lee, & Weitzner, 2008). En mettant au point les outils nécessaires à l'exploitation de ces données (algorithmes statistiques,de « *data mining* », méthodes de « *machine learning* », outils de visualisation, etc.), ces derniers ont gardé une avance importante sur toutes les autres communautés. Ces chercheurs ont ensuite rejoint des grandes structures telles que Yahoo, Google ou Facebook pour bénéficier des infrastructures et des ressources nécessaires à de telles analyses. Par exemple, Marlow, titulaire d'un doctorat au MIT Media Labs en 2001, rejoignit

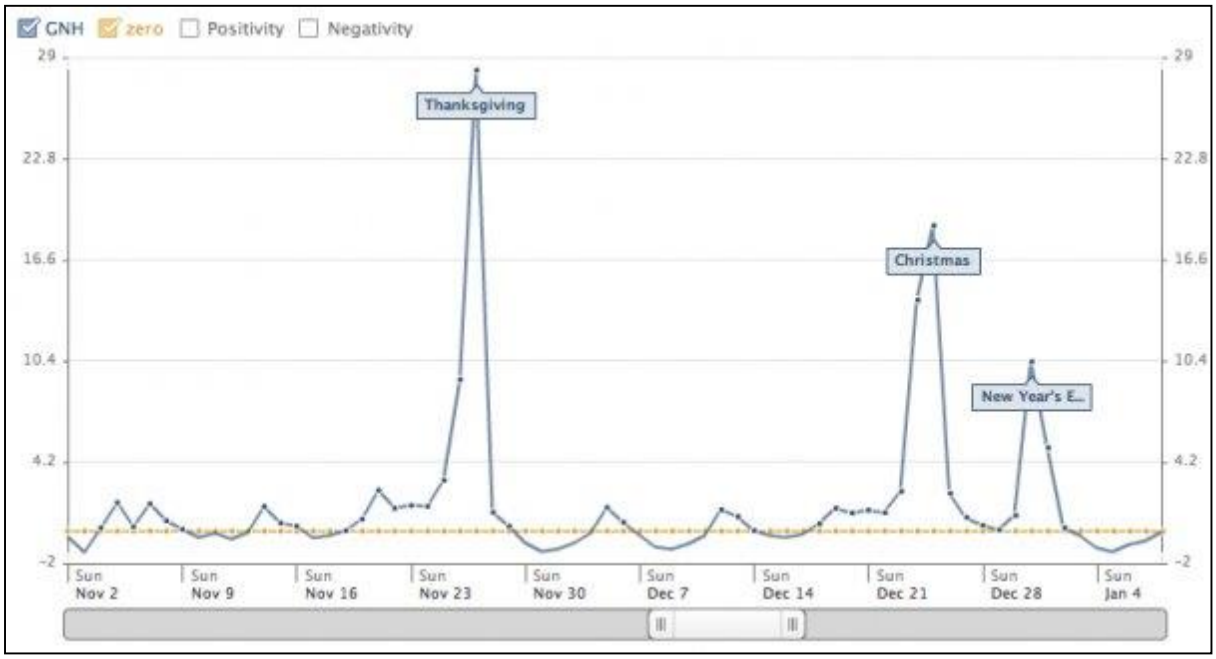


Figure 51 – Niveau de bonheur global en fonction du temps (Kramer et al. 2010)

l'équipe « *Data science* » de Facebook, connu comme le nouveau Bell Labs à l'ère des réseaux sociaux. Marlow déclara : « *pour la première fois, nous disposons d'un microscope qui nous permet, non seulement d'examiner les comportements sociaux à* un niveau de détail inédit jusqu'alors, mais également de réaliser des expériences sur des millions d'utilisateurs à la fois »[57]. Depuis, cette équipe a réalisé de nombreuses études en se basant sur la masse gigantesque de données à disposition. Elle étudia ainsi les mécanismes de contagion émotionnelle, via les réseaux sociaux (N = 289.003 ; Kramer, Guillory, & Hancock, 2014) ou encore, tenta de mesurer le niveau de bonheur (GNH) d'un pays en fonction du nombre de mots positifs ou négatifs des statuts Facebook (Figure 51) ; (N≈100.000.000 ; Kramer et al., 2010).

Ces études n'ont cependant pas pour seul objectif de générer de la connaissance pour la communauté scientifique. Elles visent avant tout à développer des moyens permettant d'influencer le comportement des utilisateurs. Le premier Mai 2012, Facebook mena une expérience visant à mesurer l'impact d'une modification des profils Facebook sur l'enregistrement en ligne en tant que nouveau donneur d'organe (Cameron et al., 2013). Facebook altéra alors sa plate-forme pour permettre aux membres de spécifier « *donneur d'organes* » dans le cadre de leur profil. Lors de ce choix, les membres étaient automatiquement orientés vers un lien leur permettant de s'enregistrer sur un site officiel, et de mettre au courant leurs amis de leur nouveau statut de donateur. Le premier jour de l'expérience, il eut 13 054 nouvelles inscriptions en ligne, ce qui représenta une augmentation de 21,1 fois par rapport à la moyenne de référence. Ces inscriptions restèrent très élevées durant les deux semaines qui suivirent. Néanmoins, il convient de préciser que la plupart des expériences que mène actuellement l'équipe *Data Science* de Facebook n'est pas aussi altruiste. La centaine de tests comportementaux réalisés, non publiés cette fois, visent plutôt à rendre plus utilisable le service en ligne ou à favoriser l'efficacité des publicités. Facebook n'est pas le seul à suivre et à étudier ses utilisateurs. Toutes les entités disposant d'une base utilisateur assez grande le font. Par exemple, *OkCupid*, sur son blog[58], avouera avoir effectué de nombreux tests sur sa clientèle, notamment pour évaluer l'efficacité réel de son système de compatibilité amoureuse (Tableau 8). On voit ainsi émerger dans ces structures (tel qu'au *New York Times* en 2014[59]) de nombreux

		number DISPLAYED to them		
		30% match	**60% match**	**90% match**
ACTUAL compatibility of users	**30% match**	10%	16%	17%
	60% match	13%	13%	16%
	90% match	16%	17%	20%

Tableau 8 – Chance de passer d'un simple message à une conversation en fonction du niveau de compatibilité affiché aux clients et leur niveau de compatibilité réel, calculé par l'algorithme OkCupid (tiré de http://blog.okcupid.com/index.php/we-experiment)

[57] http://www.technologyreview.com/featuredstory/428150/what-facebook-knows/

[58] http://blog.okcupid.com/index.php/we-experiment-on-human-beings/

[59] http://columbiaspectator.com/news/2014/02/05/math-professor-joins-new-york-times-chief-data-scientist

« Data scientist », nouveau mouton à cinq pattes, disposant de compétences en programmation, en modélisation mathématique et en marketing. Ces nouvelles *« Rock stars des IT »*[60], désignées par certains comme possédant le travail le plus sexy du XXIe siècle[61], répondent à une demande en forte croissance (à terme, 30 000 postes seront à pourvoir en France dans ce secteur[62]). Certaines entreprises même, en mal de traitement de données, organisent des concours, consistant à trouver le meilleur modèle mathématique pour un problème métier donné. Par exemple, le site « *Data Science* »[63], met en ligne tout un série de challenge orienté data, en récompensant la participation par des prix monétaire et/ou la reconnaissance d'une expertise. La « *gamification* » rend ce système particulièrement efficace pour les sociétés qui participent. En conclusion, à l'âge la captation directe des traces, dans un environnement où la beta perpétuelle pousse les ingénieurs à mettre en ligne leur système de plus en plus tôt, cette nouvelle philosophie de la conception parachève l'idée de l'obsolescence imminente du « *User Labs* » organique (Boullier, 2010a, 2012).

En conclusion, étant donné l'accaparation du cycle de conception, en amont, par les designers, et, en aval, par les scientifiques de la donnée (« *data scientist* »), où placer l'intervention ergonomique dans les IHM, compte tenu des bouleversements récents des entreprises du numérique ? La question n'est pas simple, mais mérite que l'on s'y attarde afin d'augmenter les chances de développer des méthodologies à la fois plus efficaces et en phase avec les attentes actuelles du domaine.

L'ergonome en amont de la conception, face au design

Dans le règne de la pensée créative et holistique, concentrée en début de cycle de conception, les designers se sont particulièrement démarqués ces dernières années. L'ergonomie, s'y étant intéressée plus tardivement avec l'ergonomie prospective, y a-t-elle encore un rôle à y jouer ? Il est possible de répondre à cette question en se basant sur des éléments d'ordre méthodologiques et théoriques.

D'un point de vue purement méthodologique, l'ergonomie, se basant sur la psychologie, a beaucoup à nous apprendre sur la créativité et ses mécanismes profonds. Une connaissance objective du sujet permettrait ainsi de mettre au point des méthodes stimulant la créativité de manière efficace et éprouvée. Les spécialistes sur la psychologie de la créativité ne manquent pas et de nombreuses ressources existent (Amabile, 1996; Csikszentmihalyi, 2006; Lubart, 2003). De plus, la créativité, l'activité de résolution de problèmes et l'apprentissage, trois domaines fortement développés en psychologie, peuvent être vus comme les trois facettes d'un même fonctionnement cognitif (Pochon, 2008). Ainsi, l'ergonomie possède les savoirs objectifs nécessaires pour outiller convenablement le déploiement méthodologique de la créativité.

---

60 http://www.journaldunet.com/solutions/expert/56293/les-data-scientists---nouvelles-rock-stars-de-l-it.shtml

61 https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

62 http://www.lemagit.fr/actualites/2240190718/Big-Data-des-milliers-demplois-en-perspective-aux-confins-de-la-technique-et-du-fonctionnel

63 https://www.datascience.net/fr/home/

Du côté théorique, l'ergonomie peut nourrir le champ de la conception et de l'expérience utilisateur avec une approche basée sur la psychologie positive et motivationnelle. Ce sont des approches très complémentaires, qui embrassent les besoins humains à long terme, tels que le bonheur, le bien-être ou l'accomplissement. En effet, le design nourrit actuellement la conception principalement par la stimulation hédonique et le besoin d'estime, ce qui le conduit au développement d'une expérience esthétique plutôt consumériste. Ils répondent à des besoins plus court terme et qui se renouvellent plus rapidement. Ce type d'hédonisme contemporain est décrit par Michaud comme une expérience « *sans objet* », sans accroche, où seule l'expérience de stimulation lui donne du sens. Elle rend compte d'un dynamisme qui se renouvelle sans cesse, démultipliant les expériences, le nouveau remplaçant l'anciennement nouveau. La mode se fait et se défait, créant « *des différences dans un monde où il n'y a plus de différences* » et « *s'évanouit presque aussitôt pour renaître de ses cendres l'instant d'après* » (Michaud, 2003). Ce capitalisme artiste, incorporant de manière systématique la dimension créative et imaginaire dans les secteurs de la consommation marchande, source de stimulation et d'identification à court terme pour ses utilisateurs, ne fait malheureusement pas le bonheur à long terme (Lipovetsky & Serroy, 2013). L'ergonomie a donc encore un rôle à jouer afin de rééquilibrer la conception, en étant le garant des besoins utilisateurs profonds. Guerlesquin (2012), propose ainsi un processus, basé sur une coordination des approches ergonomiques et design, autour des étapes clefs du processus de conception (Figure 52). Néanmoins, le design semble prendre conscience de ses faiblesses, et, si le basculement progressif de son fond théorique du marketing vers la psychologie se poursuit, cela tendra à combler ces lacunes. Hassenzahl fait partie de ces psychologues ayant rejoint le champ du design, il y a une dizaine d'années. Il déclara ainsi que le domaine de l'expérience utilisateur dans le design ne doit pas servir de véhicule pour le marketing, mais comme un champs d'étude permettant de concevoir des expériences ayant un sens (« *It is about creating a meaningful experience through a device* » ; Hassenzahl, 2013). Norman ajouta, en commentaire de cet article, que « *le Design se déplace d'une origine où elle s'occupait de rendre les objets attractifs (stylisme) à celui de développer des objets pouvant répondre efficacement aux besoins véritables* (…) *Chaque étape étant plus difficile que la précédente et construite sur ce qui a été appris précédemment* ».

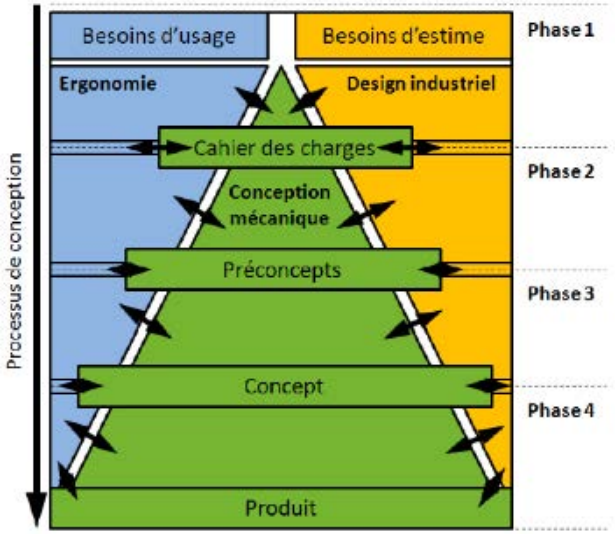


Figure 52 – Représentation du processus global de conception, alliant ergonomie et design *(Guerlesquin, 2012)*

L'ergonome face à l'évaluation intuitive, qualitative et quantitative

Une fois la phase de conception plus avancée vient la question de l'évaluation. Avec l'entrée récente de nouvelles disciplines dans le champ des IHM, de nombreuses autres épistémologies d'évaluation l'ont rejoint avec elles. De nouvelles divergences apparaissent et d'anciennes se ravivent avec l'ascendance de l'UX, d'une part, et du traitement automatisé des données, d'autre part, remettant en jeu la place, pourtant bien installée, de l'ergonomie durant cette phase. Nous verrons en quoi ces bouleversements changent la façon dont nous pouvoir l'évaluation des interfaces numériques actuellement et quels rôles l'ergonomie pourra jouer dans les années à venir.

L'évaluation experte de l'UX contre celle de la foule

Avec l'ascension de la sensibilité UX et des disciplines associées au design, d'anciens courants de pensée associés à la philosophie de l'art contribuent au débat actuel sur l'évaluation. L'expérience utilisateur, associée à l'expérience esthétique, à la recherche de l'hédonisme, à l'émotion et au beau, va dans le sens, pour un petit nombre de designer, d'une orientation évaluative associée à celle de l'art. Au contraire, les ergonomes y voient un rattachement aux modèles issus de la psychologie positive et des émotions. Cette rencontre de deux mondes, celui de la science et des arts, est bien exprimée dans un article de Norman en 2004. Il commença par y décrire une expérience personnelle qui a tourné court: celle d'avoir essayé d'expliquer sa théorie de la beauté à un historien de l'art (Norman, 2004b). Ne parlant pas le même langage, cet échange se solda par un échec. Il déclara avoir expérimenté ce que Snow désignait comme le choc des « *deux cultures* » : celle des humanités et des arts avec celle la science (C. P. Snow, 1993). Il s'agit d'un constat réalisé en 1959, mais qui est encore valable de nos jours. Ainsi, en rentrant dans le domaine des IHM, de nombreuses positions épistémologiques provenant de la philosophie des arts concurrence l'hégémonie des positions plus traditionnelles. Afin de pouvoir discuter de la place moderne de l'ergonomie dans l'évaluation des applications, il convient de prendre connaissance de ces nouveaux courants.

Parmi les points de rupture totale, la philosophie de l'art nous interroge sur le sens profond de l'évaluation, avant même de la discussion des méthodes proprement dites. En effet, si l'on postule que évaluation est une opération de nature purement subjective, il devient vain de la poursuivre au nom de tous. La question de la valeur deviendra une affaire de goût, propre à chacun. C'est la position d'Hume : « *La beauté n'est pas une qualité inhérente aux choses elles-mêmes, elle existe seulement dans l'esprit qui la contemple et chaque esprit perçoit une beauté différente* ». Dans le domaine des IHMs, ce courant s'instancie par la vision de l'artiste, exprimant sa personnalité propre dans son œuvre, et qui peut être récompensée par l'adhésion d'un groupe d'utilisateurs partageant la même sensibilité. Économiquement, ce modèle artiste peut être viable. En effet, la production en masse d'artefacts artistiques hautement différenciés est devenue rentable, car les modèles de distribution moderne permettent la mise en place d'une économie « *de la longue traîne* ». Elle se base sur une distribution d'une grande diversité de produits, chacun d'entre eux représentant une faible demande, mais qui, collectivement, représente une part de marché égale ou supérieure à celle des best-sellers.

Néanmoins, pour des produits visant à toucher une partie importante d'utilisateurs, se posera toujours la question de l'évaluation, afin de déceler le « *meilleur* » des produits pour une cible visée. De plus, l'évaluation est rendue également indispensable quand il s'agit d'estimer financièrement la valeur d'une œuvre. Pour répondre à ce problème, de nombreux philosophes analytiques affirment que les œuvres d'art peuvent être décomposées en propriétés, et que celles-ci peuvent être évaluées, pondérées, et agrégées, pour donner lieu à un score global (Dickie, 1988). Cette pratique est ancienne, et remonte probablement jusqu'au critique d'art Roger de Piles (1635-1709) qui décomposait déjà l'évaluation des peintures en quatre caractéristiques fondamentales (composition, dessin, couleurs et expression) pour ensuite les évaluer séparément sur une échelle de zéro à vingt. Cette position rejoint fortement celle de la modélisation de l'UX en IHM. Néanmoins, la position courante des historiens et philosophes de l'art de nos jours est de, soit ignorer tout simplement cette approche, soit de la critiquer très fortement[64]. Ces derniers énoncent que cet exercice est beaucoup trop hasardeux, surtout si le nombre de propriétés à juger est important et de ces dernières donnent lieu à une appréciation très subjective (style graphique, couleur, forme, …). C'est pourquoi, Hume propose, dans son essai « *sur la règle du goût* » de 1757, deux moyens supplémentaires d'évaluation : l'épreuve du temps et l'appel aux juges idéaux. Actuellement, dans le champ de l'art, c'est la méthode de l'évaluation par ces juges qui s'est imposée. C'est également celle qui est avancée par les praticiens proches de la philosophie de l'art dans le champ de l'IHM et de l'UX. Pour eux, l'estimation de la qualité d'un produit numérique ou d'une expérience d'interaction revient aux seuls experts, car la conception d'une application numérique prend en compte tellement de facteurs (utilité, utilisabilité, esthétisme, besoin d'identification, …) qu'il est impossible de l'évaluer à partir d'une modélisation objective et exhaustive de ces qualités, d'autant plus que le domaine change extrêmement vite (et est donc inaccessible par le rythme de la science). La complexité d'un tel jugement le rend également inaccessible aux novices (à comprendre : les utilisateurs), dépassés par les connaissances nécessaires à une telle tâche. Pour eux, seuls les experts du domaine (dans le meilleur des cas, les gourous du milieu, sinon les designers eux-mêmes) sont aptes à de tels jugements : Un artiste ne peut n'être évalué que par d'autres artistes. C'est un mouvement que l'on retrouve aussi dans le domaine du luxe : seuls les experts (les véritables connaisseurs) sont capables de juger de la valeur d'un Rembrandt ou d'un vin, souvent exorbitant.

Ainsi, si les experts sont les seuls ayant la compétence nécessaire, pouvons-nous faire raisonnablement confiance à leur jugement ? Les études réalisées sur les gagnants des concours pour prédire leur futur succès commercial ou la notoriété de leurs œuvres apportent une réponse négative à cette question. Par exemple, Hirschman et Pieros (1985) montrèrent que les critiques professionnelles et les récompenses sont toutes deux corrélées négativement avec la réception des auditeurs au box-office d'œuvres cinématographiques ou de Broadway. Les résultats sont également décevants pour l'étude de la postérité des films du festival de Cannes durant la période 1950 à 1970 (Ginsburgh & Weyers, 1999) ou des prix Booker des livres de fiction en Angleterre (Ginsburgh, 2003). Pourtant, l'objectif affiché par ces concours est d'aider le

[64] Pour le célèbre historien de l'art, Ernst Gombrich, la pratique de Piles est vue comme une « aberration notoire » (Gombrich, 1966, p. 66)

consommateur à séparer le bon grain de l'ivraie. James English, dans son livre, « *The Economy of Prestige »,* avança une autre raison : ces évènements conviennent surtout très bien à certains milieux, – artistes, critiques, donateurs publics ou privés, journalistes – qui « *s'emparent de et se distribuent la culture en leur donnant l'occasion de se faire remarquer, se passer l'un l'autre la main dans le dos, et se congratuler* » (English, 2005, p. 25). Ainsi, les experts se rapprochent moins des « *vrais juges* » de Hume −dotés de bon sens, de délicatesse de sentiments, d'absence de préjugés, de jugement comparatif− que des « *juges de Bourdieu* », impartis de l'autorité sociale, et dont les évaluations relèvent plus des conventions de classe (Bourdieu, 1979). D'autres études nous montrent également que la qualité de discernement de ces experts est discutable. Fritz et al. (2012) montrèrent que des musiciens professionnels ne parviennent pas à distinguer un Stradivarius d'un violon contemporain de bonne facture. De même, Ashenfelter et Quandt (1999) ont montré que l'accord entre les dégustateurs d'un même vin est faible. Mieux, Hodgson (2008) montra que ce désaccord peut apparaître aussi pour le même dégustateur, lorsqu'il goûte le même vin en des moments ou des circonstances différentes.

Les ergonomes, au contraire, se basent historiquement sur les utilisateurs pour juger de la valeur d'un produit. En effet, quand bien même ces derniers utilisent des méthodes dites « *expertes* », il s'agit de procédures fortement contrôlées et fondées sur des savoirs issus d'études utilisateur. Néanmoins, avec l'avènement du paradigme UX, un passage, partant de méthodes plutôt objectives (comme l'observation de la réussite d'une tâche) et de dimensions plus « simples » (sécurité, utilisabilité, apprenabilité, …) s'opère vers des méthodes plutôt subjectives (tel que le recueil de verbalisations ou de ressentis) et des dimensions plus insaisissable (satisfaction, émotion, immersion, …). Il ne s'agit pas ici de remettre en cause la compétence des ergonomes à relever le défi de la capture de l'UX, car, comme Barcenilla et Bastien (2009) l'ont déclaré par le passé, les ergonomes, et surtout ceux qui possèdent une approche psychologique de l'activité, détiennent les ressources théoriques et méthodologiques pour s'intégrer dans ces nouvelles orientations. Il s'agit plutôt de savoir si le recueil des impressions utilisateurs sur un produit, hautement complexe, nous permet de connaître sa valeur, hautement subjective, et, par suite, son succès futur. Car c'est bien de cela qu'il s'agit. Pour répondre à cette question, reprenons notre analyse des travaux de Galton sur la notion de sagesse des foules. Nous avons vu déjà que ce dernier avait réussi, à partir de la médiane de 787 estimations du poids d'un bœuf

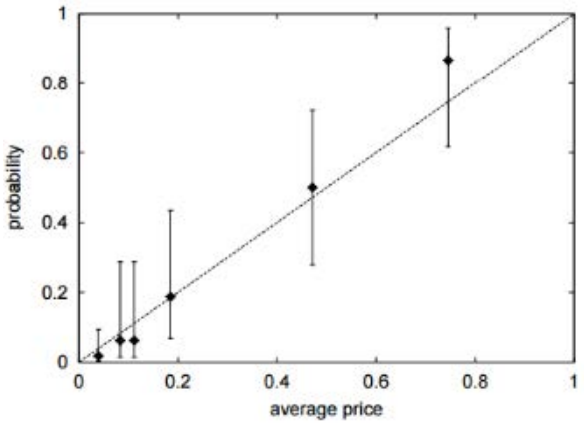


Figure 53 – Prédiction HSX sur les Oscars. Les points représente la probabilité d'être primé et la ligne pointillée la prévision parfaite (tirée de Pennock et al. *2001*)

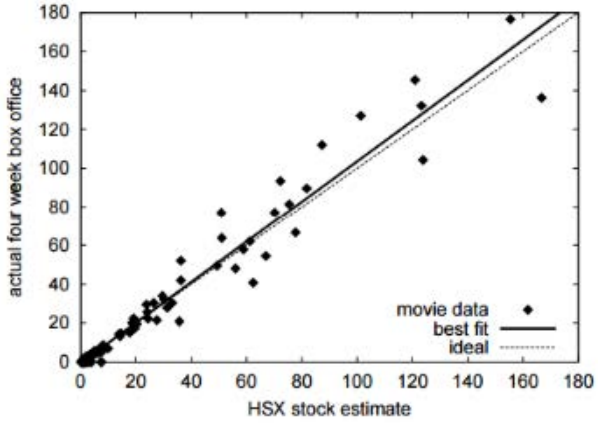


Figure 54 – Prédiction HSX sur le box-office des quatre premières semaines du film. La ligne pleine est celle de la régression linéaire et la ligne en pointillées est celle de la prévision parfaite (tirée de Pennock et al. *2001*)

de 545 kilos, d'obtenir un résultat avec moins de 1% d'erreur. Cette expérience de 1906 a été reproduite de nombreuses fois depuis, et avec, à chaque fois, des estimations étonnamment précises[65]. Dans un registre plus proche de notre sujet, c'est-à-dire l'appréciation subjective d'entités complexes, une expérience a été menée en 2007, où 57 étudiants de l'Université de Columbia ont collectivement deviné 11 sur 12 des vainqueurs aux Oscars ! (Mauboussin, 2007). Si ce résultat nous renseigne qu'un petit groupe d'individus peut faire presque aussi bien qu'un panel d'experts renommés, il existe encore mieux : le « *Hollywood Stock Exchange* [66]» (HSX). Ce simulateur de places de marché en ligne, comptant plus de 400 000 comptes utilisateurs, n'a manqué qu'un seul gagnant aux Oscars de 2004 à 2007, et prédit de manière impressionnante le résultat de film au box-office les quatre premières semaines après sa sortie (Figure 53 et 54). Ce type de jeu en ligne a, par ailleurs, été utilisé pour prédire l'issue d'élections présidentielles américaines (Berg, Nelson, & Rietz, 2008) ou de l'avancée de travaux scientifiques ou techniques (Pennock, Lawrence, Nielsen, & Giles, 2001; Pennock & Lawrence, 2001). De nombreuses leçons ont été tirées de ces études. Mauboussin (2007) constata avec l'expérience des bonbons et des Oscars plusieurs choses : une foule diverse prédit toujours mieux que des individus moyens ; la puissance prédictive d'un groupe est liée à la moyenne des connaissances individuelles et sa diversité ; et le groupe est souvent plus sage que le plus sage du groupe. Page (2007) modélisa mathématiquement ces faits, avec le théorème de la prédiction par la diversité (« *Diversity Prediction Theorem* »), qui est le suivant :

« *L'erreur **collective** (la différence entre le score vrai et l'estimation du groupe) est égale à **la moyenne des erreurs individuelles** moins **la diversité de prédiction** (la dispersion des estimations individuelles)* ».

Ainsi, si le groupe est composé, à la fois, de gens avec des connaissances non négligeables et de profils disparates, l'erreur collective devient infime. C'est pourquoi il devient très difficile de prédire mieux que le groupe, même si l'on est un expert. Ainsi, pour améliorer la clairvoyance d'un groupe, il revient au même d'améliorer la qualité moyenne des pronostics ou d'augmenter sa disparité, en introduisant des profils et des points de vue différents mais complémentaires. D'un autre côté, les profils des experts ont, certes, plus de connaissances que la moyenne, mais ont également plus tendance à se ressembler, car issus des mêmes écoles ou des mêmes cercles sociaux et professionnels. C'est pourquoi ils ne font que rarement mieux qu'un groupe conséquent et divers. Il y a cependant un corollaire à ce théorème : on constate que les problèmes les plus adaptés à la perspicacité collective sont ceux pour lesquels il n'existe pas une seule bonne méthode de résolution mais au contraire plusieurs approches, plus ou moins bonnes, que chacun explore avec plus ou moins de succès. Il sera difficile par exemple à une foule moyennement compétente en mathématiques de résoudre un problème d'une grande complexité théorique. Par contre, elle sera très compétente pour estimer la valeur d'une entité qui demande un jugement subjectif, complexe et multidimensionnel, comme cela est le cas pour l'expérience utilisateur.

---

[65] Comme l'estimation du nombre de bonbons dans un bocal, soit 850 dragibus pour une estimation collective médiane de 871 ! (Treynor, 1987)

[66] http://www.hsx.com/

Questions psychométriques associées à la dimension utilisateur de la mesure UX

La sagesse de la foule nous offre ici un parallèle intéressant avec la psychométrie. On sait que la mesure d'une entité à partir d'un instrument est toujours entachée (a) d'une part d'erreur aléatoire et (b) d'une part d'erreur systématique (Figure 55). (a) L'erreur aléatoire réduit la fiabilité de l'outil de mesure, car des variations naturelles ou non contrôlées font que les résultats fluctuent autour de la véritable valeur. C'est pourquoi, il est d'usage de multiplier les mesures, car cela permet d'annuler une partie de ces effets, la moyenne de ces mesures conduisant à faire tendre l'erreur aléatoire vers zéro. Dans le cas de la prédiction par la diversité, c'est la taille de la foule qui permet d'augmenter la fiabilité du jugement. (b) D'autre part, l'erreur systématique est vue par les psychométriciens comme un biais inhérent à l'outil et/la méthode de mesure. Il n'est pas possible de réduire l'erreur en multipliant les mesures, car toutes les mesures sont entachées de la même erreur (par exemple deux degrés de trop pour un thermomètre mal calibré). C'est pourquoi on conseille généralement d'utiliser d'autres outils en complément du premier ou de trianguler plusieurs méthodes qui ne partagent pas les mêmes biais de mesure. Ici, augmenter la diversité des méthodes, tout comme la diversité des profils de la foule, conduit à une meilleure validité de la recherche.

Nous constatons que, si nous avons l'habitude en tant qu'ergonome d'estimer la fiabilité et la validité de nos instruments de mesure, nous ne le ferons pas au sujet des utilisateurs, qui sont pourtant nos intermédiaires les plus utiles nous permettant d'évaluer un produit numérique. Pourquoi ? Pour des raisons en grande partie historiques. L'évaluation en ergonomie a hérité en grande partie des méthodes de la psychométrie, à partir desquelles les tests d'intelligence et de personnalité ont été mis au point. Il s'agissait donc de mesurer une qualité individuelle à partir d'un test dont on mesurait la fiabilité. Les contrôles de qualité psychométriques portaient donc sur les outils de mesure (questionnaires, tests, etc…) et non pas sur les individus. C'est pourquoi il est courant de calculer un alpha de Cronbach pour un questionnaire d'utilisabilité, qui est un bon indicateur de la fiabilité de l'outil. Or, actuellement, ce sont les produits numériques que nous cherchons à évaluer ! Ainsi, le contrôle du protocole d'évaluation dans les IHM doit porter à la fois sur les outils de mesure et sur les utilisateurs-évaluateurs, d'autant plus dans le domaine de l'UX. Des tests psychométriques dédiés existent, tels que la famille des coefficients de fidélité inter-juge (Von Eye & Mun, 2005). Ces outils sont très peu utilisés dans le domaine des IHM. Or, que fait un utilisateur quand on lui demande de remplir un questionnaire SUS ou AttrakDiff sinon de juger par lui-même de la qualité du produit ou de l'expérience associée ? Nous verrons qu'il existe des techniques statistiques qui permettent

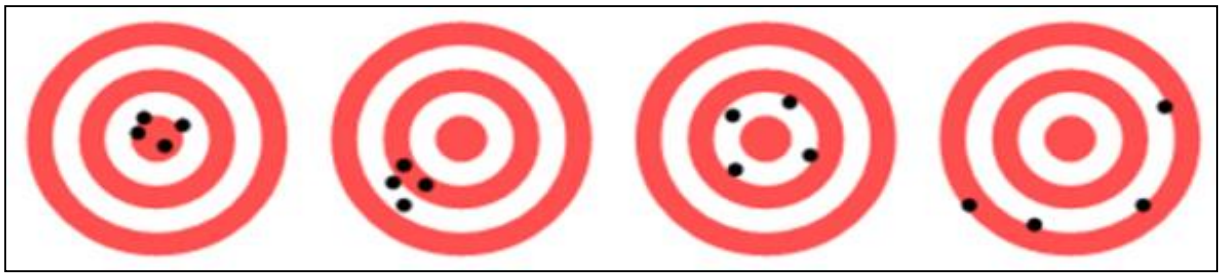


Figure 55 – Type d'erreur et ses impacts sur la validité de la mesure. De gauche à droite : (1) erreur aléatoire faible et erreur systématique faible, (2) erreur aléatoire faible et erreur systématique forte, (3) erreur aléatoire forte et erreur systématique faible, et (4) erreur aléatoire forte et erreur systématique forte

d'interroger toutes ces problématiques à la fois dans la partie triangulation (Chapitre 7). Cela est d'ailleurs un domaine d'analyse neuf dont l'ergonomie a beaucoup à apporter.

Une autre question intéressante soulevée par la facette utilisateur de la mesure est l'estimation du nombre d'utilisateurs nécessaire pour l'évaluation d'une expérience d'interaction. Dans le cadre d'évaluations sommatives, l'usage courant est que l'on interroge entre une vingtaine et une trentaine d'utilisateurs. Cette fourchette correspond au nombre minimum d'unités nécessaires pour utiliser les modèles statistiques différentiels sous-jacents à la plupart des protocoles expérimentaux classiques, plutôt qu'en relation avec l'efficacité nécessaire à l'évaluation d'un produit numérique. Dans le deuxième cas, le nombre d'utilisateurs dépendra fortement de la nature de l'indicateur que l'on cherche à estimer. Par exemple, on peut supposer que si l'on cherchera à confronter le niveau d'utilisabilité entre plusieurs interfaces, le nombre d'utilisateurs requis soit plus petit que dans le cas de la mesure de l'expérience utilisateurs associés. Bien entendu, tout dépend de ce que l'on entend par « *expérience utilisateur* », mais l'on peut raisonnablement supposer qu'un construit, liés à des dimensions motivationnelles, esthétiques ou émotionnelles (soit, plus subjectives et personnelles), crée plus de disparité d'évaluation entre les utilisateurs qu'un construit comme l'utilisabilité, beaucoup plus cognitif et objectivable. Cette évaluation demandera donc un nombre d'utilisateurs plus important, mais également plus diversifié[67].

Ce lien, entre la nature abstraite d'un construit et sa plus grande difficulté à être mesuré, a déjà été démontré dans la sphère du marketing (Cote & Buckley, 1987). Si l'on se tient à la facette « utilisateur » de l'évaluation, et non pas à la facette de l'outil d'évaluation classiquement étudiée, des travaux dans le domaine des jeux vidéo ont montré une fiabilité inter-juge faible lors d'évaluations à partir de listes d'heuristiques (Korhonen, Paavilainen, & Saarenpää, 2009; White, Mirza-babaei, McAllister, & Good, 2011)**.** Cet « *effet évaluateur* » concerne également les méthodes d'inspection plus traditionnelles et a été mis à jour qu'assez récemment (M. Hertzum & Jacobsen, 2003). Néanmoins, dans le domaine des jeux vidéo et de l'UX, une faible consistance inter-évaluateur est encore plus à redouter à cause du caractère plus abstrait et subjectif des critères d'évaluation utilisée. Le cas d'usage de l'évaluation des jeux vidéo nous a également permis de tester formellement l'hypothèse selon laquelle **l'utilisation de critères d'évaluation centrés sur l'expérience utilisateur (plus subjectif) entraîne plus de disparité de jugement entre les évaluateurs que des critères centrés sur les éléments du produit (plus objectif).** À partir d'une analyse de données issues d'une enquête réalisée sur 120 joueurs, dont certains résultats ont été publiés (Rodio & Bastien, 2013), nous avons pu montrer statistiquement que la consistance inter-évaluateur des critères d'évaluation centrés sur l'expérience du joueur était plus faible que celle des critères d'évaluation centrés sur les éléments de jeu ($n = 43$, $p = .012$), et cela, quel que soit le profil du joueur interrogé ($p < .05$, pour les joueurs amateurs et pour les joueur e-sportifs) ou le type de jeu ($p < .05$ pour les jeux de type RTS, FPS ou MMORPG). Les détails de l'étude se trouvent en annexe 1. Ces résultats sont très intéressants pour notre sujet car le domaine des jeux vidéo se tournent de plus en plus vers la mesure de l'UX. Cela nous confirme l'importance dans ce domaine de disposer de

[67] Nous pouvons raisonnablement supposer que, là où les compétences cognitives sont plus homogènes entre les individus, elle est plus disparate pour ce qui est des motivations, émotions ou sensibilité

données sur la facette « utilisateur » tant cette tâche demande une estimation importante de la part subjective du jugement.

En conclusion, nous avons vu que l'évaluation de l'expérience utilisateur est trop complexe pour être laissée dans la main de quelque expert. Au contraire, seul le nombre et la diversité des évaluateurs permet d'apprécier l'objet dans toutes ses dimensions. Néanmoins, le niveau de subjectivité de l'évaluation, qui se traduit par l'inconsistance inter-évaluateur et traduisant une sensibilité différente, nous pousse à la mesurer, afin de pouvoir mettre en œuvre des protocoles de mesures efficaces. Il s'agira donc de mobiliser des modèles statistiques permettant de prendre en compte cette facette, pourtant oubliée jusqu'à présent, dans le contrôle psychométrique en ergonomie des IHM, et qui nous permettra de mesurer sa variabilité et de décider des mesures alors à prendre. De plus, le problème du nombre et de la diversité requise pour ces évaluations nous pousse à considérer les méthodes de capture en ligne comme l'un des moyens les plus efficaces du moment pour arriver à nos fins, dans le sens où il est extrêmement plus difficile et coûteux de mobiliser un tel échantillon d'utilisateurs par des moyens conventionnels.

L'évaluation qualitative contre l'évaluation quantitative de l'UX

Deux épistémologies déchirent le champ de l'évaluation depuis ses débuts en IHM et plus encore depuis l'ascension de l'UX dans les années 2000. Nous avons déjà vu que cette confrontation est plutôt ancienne. Le même débat était présent au début des sciences sociales, et, si l'on remontre encore plus loin, depuis la confrontation des philosophes holistes et réductionnistes. C'est devenu un combat épistémologique presque mystique : les réductionnistes tendent à expliquer les phénomènes en les divisant en parties alors que les holistiques tendent à les expliquer en les réunissant. Nous en discutons encore 26 siècles plus tard, et l'enjeu est de taille, car c'est la légitimité de la division d'entités en plus petits qui justifie la pratique de la mesure, l'outil de prédilection de tout quantitativiste. Jusqu'il y a peu, c'était l'épistémologie qualitative et ses méthodes qui ont tiré leur épingle du jeu lors du passage dans le paradigme UX (Bargas-Avila & Hornbæk, 2011). Cela s'explique facilement. En effet, sans connaissances existantes au préalable et beaucoup de voies à explorer en même temps, le paradigme qualitatif est un outil idéal. Il permet une exploration rapide, heuristique, moins rigoureuse, certes, mais qui permet d'amasser rapidement des connaissances, dans toutes les directions, quand on ne part de rien. Le nombre impressionnant de méthodes UX elles-mêmes nous montrent à quel point les chercheurs ont été pressés de se donner les moyens d'explorer le champ. Toutefois, une analyse critique de ces nouvelles méthodes (ou adaptées d'anciennes) nous a montré à quel point cela s'est fait au dépend de leur qualité. Ce phénomène s'explique tout aussi facilement : dans les premières phases d'une recherche, la diversité, la souplesse et l'ouverture des méthodes permettent une exploration riche de tout l'espace de création, de ses contraintes et de ses possibilités. Et, dans un second temps, il devient nécessaire d'affiner notre jugement, pour évaluer, trancher, valider. Et c'est dans ce dernier cas que la qualité de la méthode, sa validité, sa précision, permet de donner à nos décisions notre assurance dans un milieu qui laisse de moins en moins de place à l'erreur. Dans toute activité créative, il y a un temps pour l'exploration (pensée divergente) et un temps pour la sélection (pensée convergente) ; (Guilford, 1967). Si l'état actuel des méthodes, qualitatives et créatives, permettait d'outiller convenablement les premières phases de la conception, cela n'était plus le

cas pour la suite. La raison derrière le faible nombre de recherches et de méthodes quantitatives disponible jusqu'à présent est que leurs développements demandent rigueur et patience (Law et al., 2014). De plus, la mesure de l'expérience utilisateur est un exercice particulièrement difficile, car la notion d'expérience est de nature subjective d'une part, et influencée par des circonstances contextuelles et temporelles précises d'autre part (Law, et al., 2009). Enfin, le champ des méthodes d'évaluations UX est vaste : celles-ci peuvent être produites pour des tests en laboratoire ou de terrain ; mais également à différents niveaux de détail, du feedback ressenti lors d'un clic à la relation globale se tissant entre l'usager et le produit (Obrist, et al., 2009). Le travail est donc conséquent et loin d'être terminé. Cependant, l'avancée du travail conceptuel sur l'expérience utilisateur (Obrist et al., 2011, 2012a, 2012b, 2013; Roto, Law, Vermeeren, & Hoonholt, 2011), rapproche la communauté des ergonomes d'une base assez solide pour commencer le développement de méthodes et de métriques d'évaluations sérieuses (Obrist et al., 2009; Väänänen-Vainio-Mattila et al., 2008). À ce sujet, de nombreux workshops et rassemblements scientifiques ont été organisés dans ce but, mais le panel des méthodes d'évaluation dans la sphère de l'expérience utilisateur demeure incomplet (Obrist, et al., 2009). Malgré cela, on constate une nette augmentation des études sur la mesure de l'UX depuis les années 2000. En effet, pour Google Scholar, le nombre d'articles retournés avec les mots-clés « *User Experience Measures* » et « *Measure User Experience* » était respectivement de 37 et 134 dans les années 2000, contre seulement 0 et 3 dans les années 90 (Law, 2011). Cela a poussé dernièrement les auteurs à mener des enquêtes approfondies sur le rôle et la visée épistémologique de la mesure UX dans le champ actuel des IHM (Law & Abrahão, 2014; Law et al., 2014; Law, 2011). Il est donc possible aujourd'hui de cerner deux camps, bien équilibrés et aux positions bien affirmées. Il s'agit ici d'exposer leurs argumentations et de présenter l'avenir de la mesure dans l'UX, qui est une voie de recherche prometteuse pour toute l'ergonomie des IHM.

Pour Law (2011), il y a les « *Chercheurs UX basés sur le Design* » et les « *Chercheurs UX basés sur les modèles* ». Les premiers se concentrent sur l'articulation des expériences dans toutes leurs richesses et leurs contextes, alors que les deuxièmes se concentrent sur l'étude de l'expérience utilisateur dans l'optique de permettre sa généralisabilité et favoriser leurs comparaisons (Boehner, Depaula, Dourish, & Sengers, 2007). Law (2011) cite, comme exemple de chercheurs dans le premier camp, Blythe, Cockton, Forlizzi, Gaver, McCarthy, Monk et Wright, et, dans le second camp, Hassenzahl, Mahlke, Sutcliffe, Tractinsky, et van Schaik (nous pouvons rajouter également Law), bien que, toujours selon Law (2011), des migrations existent entre les deux camps, particulièrement ces derniers temps, du premier vers le second. Chacun des deux camps mettent en avant leurs positions respectives sur la pertinence de la mesure UX, en puisant dans leurs héritages théoriques propres : les premiers dans l'ingénierie[68] et la psychométrie, où la mesure fait partie des méthodes indispensables à la construction de la connaissance ; les deuxièmes dans les « *humanités* » où la mesure est considérée souvent comme simpliste et naïve, surtout quand l'entité à mesurer est mal-définie et/ou trop sujette à l'interprétation (Bartholomew, 2006). Pour ce dernier point, il convient de

[68] On citera William Thomson et son célèbre : « *mesurer c'est connaitre* » pour montrer l'attachement académique des ingénieurs à la nécessité de la mesure

préciser qu'il est fort bien reçu dans les deux camps, étant donné que le travail de définition et d'opérationalisation (mise en mesure) d'un concept se réalise en tandem : il est impossible de mesurer quelque chose avant de savoir précisément ce qu'on mesure. Le premier camp, souvent composé de professionnels ayant eu une formation ou un emploi centré sur le design (Kaye et al., 2011; Roto, Law, Vermeeren, & Hoonhout, 2010), avance une série d'arguments démontrant leurs scepticisme envers la mesure de l'UX. Trois sont particulièrement redondants : il s'agit (i) de l'incapacité de la mesure à cerner les construits complexes, (ii) le faible pouvoir informationnel des méthodes quantitatives pour la conception et (iii) la quantité de ressources nécessaires trop importantes pour ce type d'évaluation.

(i) Le premier argument est illustré par le commentaire de Forlizzi et Battarbee (2004) : « *les réponses émotionnelles sont difficiles à comprendre et encore plus à quantifier* » (p. 265). C'est la remarque la plus avancée dans l'enquête de Law et al. (2014) contre la mesure de l'UX : « *Une expérience n'est pas mesurable de la même façon que pour les distances ou le poids. Nous devons nous baser sur les interprétations subjectives* » (P319). Contre cet argument, Hassenzahl (2008) avance que l'unicité et la variabilité des expériences avec la technologie est bien moindre que celle impliquée dans le cadre d'autres approches phénoménologiques. De plus, c'est sans compter sur l'expertise des psychologues, dont l'histoire empirique a toujours été de rendre mesurable ce que la plupart des gens ordinaires pensaient immesurables, tels que les cognitions, les émotions, les motivations, les traits de personnalité et bien plus encore (Eid & Diener, 2006). Pour le moment, il est vrai que beaucoup de mesures et de méthodes sont tirées du précédent paradigme, c'est-à-dire de l'utilisabilité (Tullis & Albert, 2008), et que l'UX est un concept qui mélange de nombreux facteurs psychologiques, sociologiques et physiologiques. Cependant, nous ne partons pas de zéro. Du côté des recherches sur la mesure des émotions par exemple, qui est un des sujets majeurs dans le domaine UX (McCarthy & Wright, 2004), nous pouvons nous appuyer sur plus d'un siècle de travail, de la théorie des émotions de James-Lange (dont une synthèse du travail a été réalisée par Lang , 1994), aux travaux célèbres d'Ekman, Russell, Frijda, ou encore Scherer (Ekman, 1992; Frijda, 1986; Russell, 2009; Scherer, 1989). Par ailleurs, leurs applications dans le contexte de l'UX ont montré que la mesure des émotions était plausible, utile et même nécessaire (Bargas-Avila & Hornbæk, 2011; Coan & Allen, 2007). Néanmoins, il convient de rappeler que toutes les mesures psychologiques ne sont que des approximatisations de phénomènes réels plus complexes, et doivent donc toujours être considérées de manière critique (Hand, 2004).

(ii) Concernant le deuxième argument avancé contre la mesure de l'UX, c'est-à-dire son faible pouvoir informationnel, les déclarations de Swallow et al. (2005) en sont un exemple : « *ces approches sont utiles dans le cadre expérimental, mais peuvent nous faire passer à côté de certains insights qui résistent mal à de telles réductions (…) les données qualitatives possédant un niveau de richesse et de détails qui peut être absent des mesures quantitatives* » (p. 91-92). Ce point de vue résonne également dans les propos de Höök (Roto et al., 2010) : « *la question est de savoir si l'évaluation de l'expérience de l'utilisateur final à partir de mesures simplistes nous aide vraiment à produire de meilleurs designs ou à mieux comprendre l'expérience utilisateur. De mon point de vue, il y a trop de réductionnistes ici qui blessent la recherche dans le domaine en prétendant fournir des mesures et des méthodes qui permettront à tout le monde d'évaluer la 'valeur' UX d'un système* » (p. 17). Ces propos montrent de nouveau la confusion

de certains acteurs entre les avantages et bénéfices à faire valoir en fonction des méthodes. Les méthodes qualitatives n'ont jamais revendiqué un pouvoir de diagnostic aussi élevé que les méthodes quantitatives. Elles n'ont tout simplement pas le même but, et nous l'avons répété de nombreuses fois dans cette thèse. Les évaluations formatives –qualitatives– excellent dans les premières phases de conception, car elles permettent un diagnostic riche des choix de conception déployés ; mais ce sont les méthodes quantitatives qui permettent d'estimer avec fiabilité la valeur d'un produit ou de trancher entre plusieurs designs de manière plus définitive. L'enquête menée par Law et al. (2014), va dans ce sens, et les arguments recueillis en faveur de la mesure de l'UX pointent en premier sa valeur pour l'évaluation sommative, c'est-à-dire de pouvoir justifier les décisions prises, valider les objectifs de conception, et, cela, avec une fiabilité suffisante : « si *nous utilisons que l'intuition du designer, seulement son interprétation emphatique, cela manque de fiabilité pour le reste du monde* » (S2, p.531).

(iii) Seulement, et nous en venons au troisième argument contre les mesures, les mesures UX sont vues comme particulièrement coûteuses. Dans l'enquête de Law et al. (2014), elles sont même perçues comme douloureuses… psychologiquement, que ce soit au moment de la définition du protocole, le recrutement des participants (plus nombreux), de la passation méthodologiquement plus lourde ou du traitement statistique des données plus périlleux. « Il est très coûteux de mettre en place des évaluations fiables, car cela requiert des mois de préparation, et cela est dur à justifier (P290 ; Law et al., 2014). C'est, de mon point de vue, la limite la plus importante des méthodes quantitatives actuelles. En effet, avec l'accélération du développement technologique et du renouvellement des produits numériques associés, être réactif sur le marché est devenu une qualité vitale. C'est pour cela qu'il est important de faire évaluer les méthodes quantitatives pour qu'elles restent utilisables. L'une de ces voies a été ouverte par les progrès du numérique : son automatisation. Je pense que, dans ce domaine particulier, l'ergonomie a un rôle majeur à jouer, bien que cela soit les ingénieurs informaticiens qui en tiennent pour le moment les rênes.

L'ergonome face aux data scientist pour l'évaluation automatisée

Les limites de l'approche algorithmique pure

Nous avons vu dans les chapitres précédents que l'évaluation automatisée, grâce au Big Data et la capture directe des traces, est devenue l'apanage des ingénieurs informaticiens, et plus précisément des « *Data Scientist* ». En effet, quoi de plus logique pour les experts du numérique que de s'accaparer l'avenir des méthodes quantitatives , car, par définition, la mesure, comme le numérique, a pour objectif la « *mise en nombre* » (Hand, 2004; Taillet, Villain, & Febvre, 2013). Reste-t-il donc encore un rôle à jouer dans ce domaine pour les ergonomes ? Pour cela, revenons de plus près sur l'expérience de Facebook sur la mesure du niveau de bonheur (GNH) d'un pays en fonction du nombre de mots positifs ou négatifs, tirés à partir des statuts Facebook (Figure 51) ; (N≈100.000.000 ; Kramer et al., 2010). Cette nouvelle mesure, le FGNH (« *Facebook's Gross National Happiness »*), avait démontré une certaine validité, par la présence d'une variation cyclique en fonction des jours de la semaine et d'une augmentation perceptible

lors des vacances nationales. Afin de tester plus finement la validité de cette mesure, une équipe de psychologues et de psychométriciens décida de comparer les résultats du FGNH avec une échelle du bonheur déjà validée, la SWLS (« *Satisfaction with Life Scale* »). Ils ne trouvèrent pas de corrélation significative entre les deux, bien que l'échelle SWLS montrât toutefois une relation positive avec le nombre de mots négatifs recueillis à partir des statuts Facebook (Wang, Kosinski, Stillwell, & Rust, 2014). Ainsi, ils conclurent que le FGNH n'était ni une mesure valide de l'humeur, ni du bien-être ; mais que, peut-être, jouait-elle un rôle dans la régulation de l'humeur. Quelle leçon devrait tirer les « *Data Scientist* » de Facebook d'une telle étude ? Que la validité faciale d'une mesure n'est pas suffisante pour estimer sa pertinence et que d'autres tests doivent être réalisés, tel que celui de la validité concourante réalisé par Wang et al. (2014). Ainsi, si la bonne utilisation des méthodes de capture, via des outils de datamining (et dans une certaine limite le traitement statistique associé) n'est généralement pas remise en cause, d'autres aspects, en relation avec les méthodologies de recherche ou le bagage théorique nécessaire (ici la conception du bonheur) pour mener à bien ce genre d'étude, sont souvent lacunaires. Or, l'âge de la capture des traces entraîne une mutation d'importance qui devrait générer autant de précautions que lors des précédents âges (Boullier, 2012). Or, certaines caractéristiques épistémologiques, voire culturelles des ingénieurs, couplées aux moyens extraordinaires d'accès aux données, les font passer (volontairement ou involontairement) à côté de certaines précautions méthodologiques, telles que (i) l'échantillonnage, (ii) la méthode hypothético-déductive ou (iii) la modélisation/théorisation.

(i) Dans l'Age du Big Data, où nous observons toutes les actions utilisateurs, en toutes heures, n'observons-nous pas simplement **tout** ? Les notions d'échantillons et de population a-t-elle encore lieu d'exister ? C'est en tout cas ce que déclarent les partisans du nouveau paradigme N=ALL (Mayer-Schönberger & Cukier, 2013). Toutefois, cela est plus une supposition sur les données qu'un fait avéré (Fung, 2013). La réalité est autre : nous n'observons jamais tout, et, dans un certain nombre de cas, les données auxquelles nous avons accès automatiquement ne nous permettent justement pas de répondre correctement à nos interrogations (Schutt & O'Neil, 2014). En effet, même si nous avions accès à tout le corpus de Twitter, Google et Facebook, étendre une inférence obtenue à partir de ces données à tout le fonctionnement humain est un saut bien périlleux à franchir. C'est ce qui fait qu'une personne d'âge moyen trouvera les recommandations de Netflix peu pertinentes pour elle, car les majorités des personnes qui prennent la peine de noter les séries sont des personnes jeunes, avec des goûts donc propres à leur âge. Dans son strata talk « *les biais cachés du Big Data* », Kate Crawford donna un autre exemple basé sur l'analyse des tweets avant et après la tornade Sandy (Crawford, 2013). Quand on analyse les tweets émis avant et après Sandy, on a l'impression

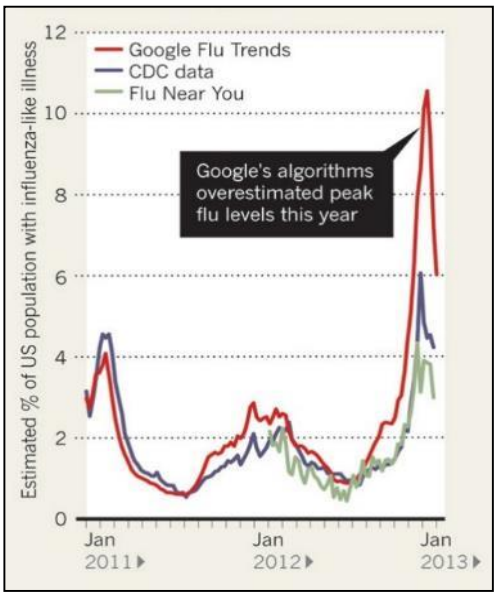


Figure 56 – Comparaison de trois méthodes de mesure des épidémies de grippe au Etats-Unis (tiré de *Butler, 2003*)

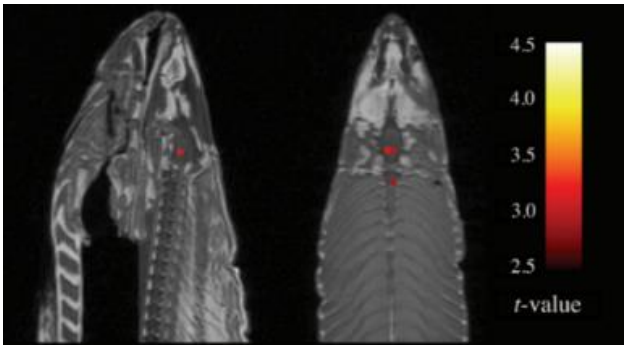


Figure 57 – Régions du cerveau présentant un signal BOLD significatif (tiré de *Bennett et al, 2010*)

que la plupart des gens allaient faire leurs courses avant la tornade pour la fêter juste après… alors qu'en fait, la majeure partie des tweets était issue de Newyorkais, et que les côtiers du New Jersey, qui se préoccupaient plutôt de leurs maisons qui s'écroulaient, avait mieux à faire que tweeter. Ainsi, malgré le nombre important de données récoltées, les milliers de tweets ne suffirent pas à constituer un échantillon représentatif de la situation. Un autre exemple célèbre (qui ne date pas d'hier) nous est donné par Harford (2014) : il relate la célèbre prédiction ratée d'Alf Landon lors des présidentielles américaines de 1936. Elle était basée sur une enquête téléphonique de plus de deux millions de personnes. Mais ce que le magazine n'avait pas réalisé, c'est que ce segment de la population durant la grande dépression, c'est-à-dire possédant un téléphone et une voiture, n'était pas représentatif de la population générale. À l'inverse, le sondage d'opinion de George Gallup, qui se basait sur un échantillon beaucoup plus petit mais plus équilibré, prédit correctement l'élection de Franklin Roosevelt. Ainsi, dans une étude, la quantité des données ne suffit pas à prédire sa qualité.

(ii) Une autre tentation des ingénieurs du Big Data a été de s'affranchir de la nécessité de déceler les liens de causalité, en exploitant à la place les liens de corrélation entre des millions de données. Cukier et Mayer-Schoenberger (2013) déclarèrent ainsi que la révolution du Big Data consiste, à la fois, en la collection d'un grand ensemble de données, l'acceptation de leurs désordres et l'abandon de la recherche des causes. Pour eux, le Big data n'a pas besoin de comprendre les causes, car les données récoltées sont assez importantes pour que cela ne soit plus nécessaire. De même, il n'y a pas lieu de s'inquiéter de l'erreur d'échantillonnage, car la masse de données permettra littéralement de « *suivre à la trace la vérité* » (Cukier & Mayer-Schoenberger, 2013). L'exemple prestigieux souvent cité est celui de l'indicateur GFT (Google Flu Trend; Ginsberg et al., 2009) qui, à partir de l'analyse de milliers de requêtes de recherche Google, était capable de prédire très précisément les épidémies de grippe…jusqu'en hiver 2013. En effet, à ce moment, l'indice commença à surévaluer de plus de deux fois (Figure 56) la propagation du virus sans que l'on sache pourquoi (Butler, 2013). En effet, le problème avec ce modèle riche en données mais pauvre en théorie, c'est qu'il n'y a pas moyen de savoir ce qui liait, à la base, les requêtes de recherche Google avec les épidémies de grippe. En fait, on n'a jamais cherché à le savoir, en essayant de remontrer aux causes. Google s'est simplement appuyé, par effet d'aubaine, sur des patterns corrélationnels intéressants à un moment T de l'histoire. Le problème est qu'une corrélation sans théories explicatives est inévitablement fragile : sans connaître la/les cause(s) derrière la corrélation, impossible de savoir ce qui la brise. Cela est sans compter sur un piège encore plus pervers de la corrélation, le fameux sophisme « *Cum hoc ergo propter hoc* » (du latin : « *Avec ceci, donc cause de ceci* »), qui fait confondre corrélation et causalité. En effet, si l'on entraîne un algorithme de recrutement sur des données passées, et qu'il se rend compte que les femmes tendent à démissionner plus souvent, sont moins promues et reçoivent un feedback en général plus négatif que les hommes,

on serait tenté de recruter des hommes à compétences égales…sans prendre en compte la possibilité que c'est peut-être l'entreprise elle-même qui traite mal les femmes (Schutt & O'Neil, 2014). Enfin, le nombre de variables étudiés lors des études corrélationnelles peut également être un piège. C'est ce que nous apprend de manière ludique l'expérience de Bennett et al. (2010) avec l'étude d'un saumon mort. Ils mirent le saumon dans un scanner fMRI et lui montrèrent une série de photographies représentant des personnes humaines dans différentes situations sociales. Puis, on demanda au saumon de déterminer quelles émotions les personnes devaient ressentir. En testant toutes les régions du cerveau du saumon, les chercheurs trouvèrent une activation neurale significative dans certaines zones entre la situation de repos et celle de présentations des images ($t(131) > 3.15$, $p < .001$). Cela montre qu'en effectuant un grand nombre de test, il est possible de trouver des liens même dans des conditions où leur absence est certaine. À l'époque de l'étude, entre 25 et 40% des recherches utilisant la technologie fMRI ne prenait pas les précautions nécessaires méthodologiques pour contrer ce type de biais (C. M. Bennett et al., 2010). Le problème est que les partisans de la recherche automatique de connaissance, via le *data mining*, par l'exploration de masse énorme de données, se retrouvent confrontés, sans le savoir, à ce type de biais. Si l'on teste aveuglément, tous les liens possibles dans une masse énorme de données, on finira TOUJOURS par trouver quelque chose, même dans des données complètement aléatoires : « *si vous torturez des données assez longtemps, elle vous confesseront ce que vous voudrez* » (Hal Varian, cité de Tullock, 2001). On appelle cette pratique le «*data dredging*», le «*data fishing*», «*data snooping*», «*equation fitting*» ou encore le «*p-hacking*» (Jensen, 2000; Selvin & Stuart, 1966; White, 2000). Le site « *Spurious Correlations* » en a fait sa spécialité et publie diverses corrélations saugrenues, dont une corrélation entre le budget américain en science, aérospatial et technologie et le nombre de suicides par pendaison, étranglement ou suffocation[69], ou encore un lien entre le taux de divorce dans le Maine et l'évolution de la consommation de margarine aux États-Unis[70].

(iii) Cela nous mène à la troisième tentation des chercheurs informatiques en Big Data, celui de s'affranchir tout simplement des modèles théoriques, voire de la méthode scientifique elle-même. Selon la logique développée, puisque la totalité de données existantes seront récupérées, les anciens modèles statistiques et scientifiques deviennent obsolètes. L'éditeur en chef de Wired, dans un article de 2008 au titre provocateur de « *La Fin de la théorie : le déluge de données rend la méthode scientifique obsolète* », va encore plus loin : pour lui, avec assez de données, les nombres parlent d'eux-mêmes. Dans le même ordre d'idées, l'utilisation d'approches massivement statistiques, à la place des théories linguistiques dans le domaine de la linguistique computationnelle, fit dire à Yehoshua Bar-Hillel et Yorik Willks que « *ce sera un "sac d'astuces" et non la théorie qui fera avancer la linguistique computationnelle dans le futur* » (Wilkes, 1996). Cette manière de se focaliser sur les techniques mathématiques et algorithmiques, en balayant toute autre considération théorique et méthodologique, illustre bien la loi du marteau de Maslow (1964) qui nous pousse à transformer la réalité d'un problème en

[69] Spurious Correlations n° 1597. US spending on science, space, and technology correlates with Suicides by hanging, strangulation and suffocation, disponible ici : http://www.tylervigen.com/view_correlation?id=1597

[70] Spurious Correlations n° 1703. Divorce rate in Maine correlates with Per capita consumption of margarine (US), disponible ici : http://www.tylervigen.com/view_correlation?id=1703

fonction des réponses que l'on dispose. L'effort et le temps consacré à l'apprentissage de compétences de recherches particulières bornent le chercheur à un nombre limité de tactiques et le poussent à devenir « *aveugle* » à d'autres. Hetherington (2000) constate ainsi que les livres de statistiques multivariés de ces 25 dernières années ne fournissent qu'une base sommaire du rôle des théories dans l'entreprise scientifique. Cela a encore été exacerbé par la plus grande puissance computationnelle disponible et la sophistication des techniques d'inférences statistiques. Comme Cattell l'observait déjà en 1988 : « *hélas , il faut aussi reconnaître que l'ordinateur a apporté le danger de magnifier les erreurs de conception , de propager des abus vulgaires , et de favoriser la collecte des données non-inspirées* » (Cattell, 1988, p. 9). Pour Hetherington (2000), l'utilisation de modèle d'apprentissage informatique automatisé ne peut pas se passer des théories qui permettent de (a) sélectionner les variables à intégrer dans le modèle, (b) spécifier leurs relations, et (c) interpréter les résultats obtenus. (a) L'adage informatique célèbre « *Garbage In, Garbage Out* » (ordures à l'entrée, ordures à la sortie) ou GIGO reste opérationnel. Même avec le modèle statistique le plus puissant, si on le nourrit en entrée de mauvaises données, ce que l'on obtient en sortie le sera nécessairement aussi. La théorie joue un rôle essentiel dans la sélection des variables pour la recherche, car elle précise non seulement ce qui **doit** être observé, mais également **comment** l'observer. Au moins deux types d'erreurs (Simon & Newell, 1963) peuvent se produire lors de la modélisation des phénomènes complexes: l'exclusion de variables pertinentes (erreur d'omission) et de l'inclusion de variables non pertinentes (erreur de commission), les deux ayant un effet dévastateur sur la qualité du modèle (Pedhazur, 1982). De même, l'utilisation d'un grand nombre de variables injectés aveuglement dans le modèle est également insoutenable pour le modèle d'un point de vue purement statistique, car la taille du jeu de données nécessaires grandit exponentiellement en fonction du nombre de variables injectées dans le modèle. C'est-à-dire que plus le modèle est complexe et automatisé, plus il demande de données en entrée. Ce phénomène fut appelé par Richard Bellman (1957) la « *Malédiction de la dimension* » (« *Curse of dimensionality* ») et concerne particulièrement l'apprentissage automatique et la fouille de données. Dans ce cadre, la théorie permet d'identifier un lot parcimonieux de variables pertinentes, contrairement à l'approche inefficace de la modélisation aveugle. (b) La théorie est également essentielle pour préciser exactement quel type de relations (hiérarchiques, associatives ou causales) existe ou n'existe pas entre les variables et leurs concepts sous-jacents. En conséquence, la théorie a des implications précises pour la définition opérationnelle des construits, le choix du modèle mathématique utilisé, et enfin l'évaluation de sa validité. (c) Dernièrement, la théorie joue également un rôle essentiel dans l'interprétation des résultats obtenus. La théorie est le schéma primaire, ou cadre de référence par lequel le chercheur comprend le contenu et les implications des résultats de la recherche. Balayer cela de la recherche laissera le chercheur au milieu d'un océan de faits inintelligibles ou le fera succomber à des résultats illusoires (Cliff, 1983).

Ainsi, la question théorique majeure concernant la mesure automatisée de l'expérience utilisateur concerne directement sa définition. Comment mesurer les émotions, l'immersion ou encore l'engagement d'un utilisateur si l'on ne saisit pas précisément ce que ces notions représentent ? Au DataEDGE, une conférence tenue chaque année à l'école de l'Université de Michael Chui (du McKinsey Global Institute) et Itamar Rosenn (premier Data scientist de

Facebook) discutèrent ainsi des difficultés à définir un « *utilisateur engagé* » (Schutt & O'Neil, 2014). Cela est une question qui ne va pas de soi. Or, dans la plupart des papiers concernant l'évaluation de construits latents complexes par le *machine learning*, beaucoup d'efforts méthodologiques sont investis dans la construction du modèle mathématique et peu dans ce qui est mesuré. Or, la **précision d'une mesure doit se référer constamment à la maturité conceptuelle de l'objet mesuré** : « *Mesurer exactement un objet fuyant ou indéterminé, mesurer exactement un objet fixe et bien déterminé avec un instrument grossier, voilà deux types d'occupations vaines que rejette de prime abord la discipline scientifique. »* (Bachelard, 1970, p. 213). Disposer d'un modèle mathématique complexe n'est pas le garant d'une mesure de grande qualité si l'on ne prend pas la peine de se pencher tout autant sur ce que l'on veut mesurer. En effet, la qualité d'une mesure est comme une chaîne dont la valeur est liée à son élément le plus fragile. Rien ne sert d'avoir une mesure calculée par une méthode complexe à 15 chiffres après la virgule, si l'on se base sur un recueil des données en amont d'une précision inférieure : *« une précision sur un résultat, quand elle dépasse la précision sur les données expérimentales, est très exactement la détermination du néant. Les décimales du calcul n'appartiennent pas à l'objet (...) Cette pratique rappelle la plaisanterie de Dulong qui disait d'un expérimentateur : il est sûr du troisième chiffre après la virgule, c'est sur le premier qu'il hésite »* (Bachelard, 1970, pp. 214,241–242). Or, les systèmes de mesures automatiques présentés par les ingénieurs informaticiens sont en général des boites noires faussement objectives (Boullier & Lohard, 2012) : « *le packaging des offres réalisé par le marketing, porté par une demande considérable de suivi de ce continent improbable qu'est le web 2.0, permet de produire des boîtes noires dont les composantes ne sont jamais décrites. La critique est alors compréhensible lorsque des annonces spectaculaires sont faites à travers des dashboards (tableaux de bord) tous aussi affinés et pilotables à souhait apparemment, alors qu'en réalité, les algorithmes ont éliminé tous les énoncés problématiques et toutes les sources incertaines* ». Elles tendent, comme dans les sciences économiques, à nier les facteurs humains par un processus de naturalisation. Ils finissent ainsi par présenter des modèles, sans parfois s'en rendre compte, partielles et partiales. C'est une vision épistémologique en complète opposition avec les designers abreuvés de philosophie de l'art. Là où la position des ingénieurs Big Data est de s'émanciper des sciences molles par la prétention une objectivité positiviste pure, le rêve des designers (ou tout du moins des philosophes artistes) est de s'émanciper des sciences dures (herméneutique, courant postmodernisme, relativiste) par la subjectivité pure. Or, la vérité se trouvant souvent au milieu, la psychologie est bien équipée méthodologiquement et épistémologiquement pour permettre l'appréciation d'objets complexes sans tomber dans aucun de ces excès.

La reconnaissance du social data scientist

Dernièrement, de nombreux data scientist reconnaissent certaines erreurs de jeunesse et essayent d'y remédier. Parmi les erreurs les plus communes, William Chen cite le fait de penser que le domaine du *data science* n'est affaire que de mathématique et d'informatique. Or, il est avant tout question de statistique et, plus encore, du produit : c'est-à-dire qu'il s'agit de quantifier ce qui fait qu'il est *« bon* », de mesurer sa « *qualité* » et l' « *engagement* » des utilisateurs (Shan, Chen, Wang, & Song, 2015). Amit Moran, le Chief Data Scientist chez

Crosswise[71], cite l'erreur commune de la focalisation algorithmique au lieu de se concentrer sur les données alors que le choix de l'algorithme est souvent mineur comparé à la sélection des bonnes caractéristiques[72] et l'évitement de certains biais. Enfin, Franscesca Milletti, Data scientist chez Roche[73], insiste sur l'importance de l'expertise dans le domaine d'étude, largement sous-estimé par la plupart des data scientist. Ainsi, même si on entend dire souvent qu'il faut laisser « *les données parler d'elles-mêmes* », encore faut-il poser les bonnes questions et sous la bonne forme, en s'appuyant sur les connaissances préexistantes du domaine. Pour elle, un data scientist sans la connaissance du domaine est comme un corps sans esprit. Cela fait écho au diagramme de Venn de Drew Conway[74], qui définit l'expertise du domaine comme un des trois piliers de la *data science*, aux côtés des compétences en hacking, statistiques et mathématiques (Figure 58). Ce qui fait dire à Schutt et O'Neil (2014) que, puisqu'un même individu ne peut pas maitriser toutes ces disciplines en même temps, une équipe constituée de profils différents mais complémentaires, dans l'objectif de résoudre un problème particulier, est nécessaire (Figure 59). Ainsi, de nouvelles équipes de *data scientist*, composées de profils classiques ou atypiques, investissent des champs d'études existants avec des outils nouveaux. Elles contribuent sans peine au développement de champs scientifiques durs ou du vivant (physique, biologie, …), de la médecine, comme l'e-santé (Atallah & Yang, 2009; Lorenz & Oppermann, 2009; Wicks, Vaughan, Massagli, & Heywood, 2011), ou encore des humanités numériques (Rieder & Röhle, 2012; Rogers, 2015), tel que la *Culturomics* (Michel, Shen, Aiden, Veres, & Gray, 2011) ou la sociologie computationnelle (Varenne, 2011). D'ailleurs, de nombreux *data scientist*, voulant démontrer la pertinence de l'étude de la sociologie avec des méthodes « *data-intensives* », citent les travaux de Bruno Latour, qui, lui-même, évoque l'idée de Gabriel Tarde selon laquelle les sciences sociales sont en réalité plus quantitatives que les sciences naturelles. « *Quoi de plus rafraîchissant dans Tarde (un siècle plus tard !) qui n'avait jamais douté une minute qu'il serait possible d'avoir une sociologie scientifique – ou plutôt, selon ses termes, d'une interpsychologie* » (Latour, 2009, pp. 147–148).

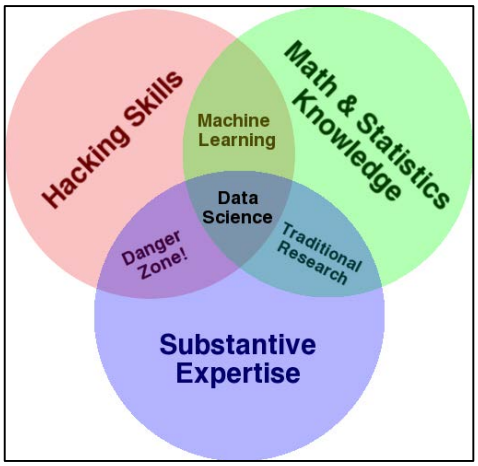


Figure 58 – Diagramme de Venn de Drew Conway des compétences nécessaires dans le domaine des Data Science[74]

Steve Lohr[75] déclara que « *le Big Data est un terme vague qui est utilisé un peu trop souvent ces derniers temps. Mais pour faire simple, cette notion accrocheuse signifie trois choses. Premièrement, c'est un pack de technologies. Deuxièmement, c'est une révolution potentielle de la mesure. Troisièmement, c'est un point de vue, une philosophie, de comment les décisions seront -ou peut-être devront- être prises dans le futur* ». L'ergonomie, et par son bras armé la psychométrie, peut potentiellement envisager, grâce à ces nouveaux outils et le sérieux de ses

[71] https://angel.co/crosswise

[72] Dans le *machine learning*, les caractéristiques (ou « *features* » en anglais) sont les propriétés mesurables d'un phénomène observé

[73] http://www.roche.com/index.htm

[74] http://www.dataists.com/2010/09/the-data-science-venn-diagram/

[75] http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

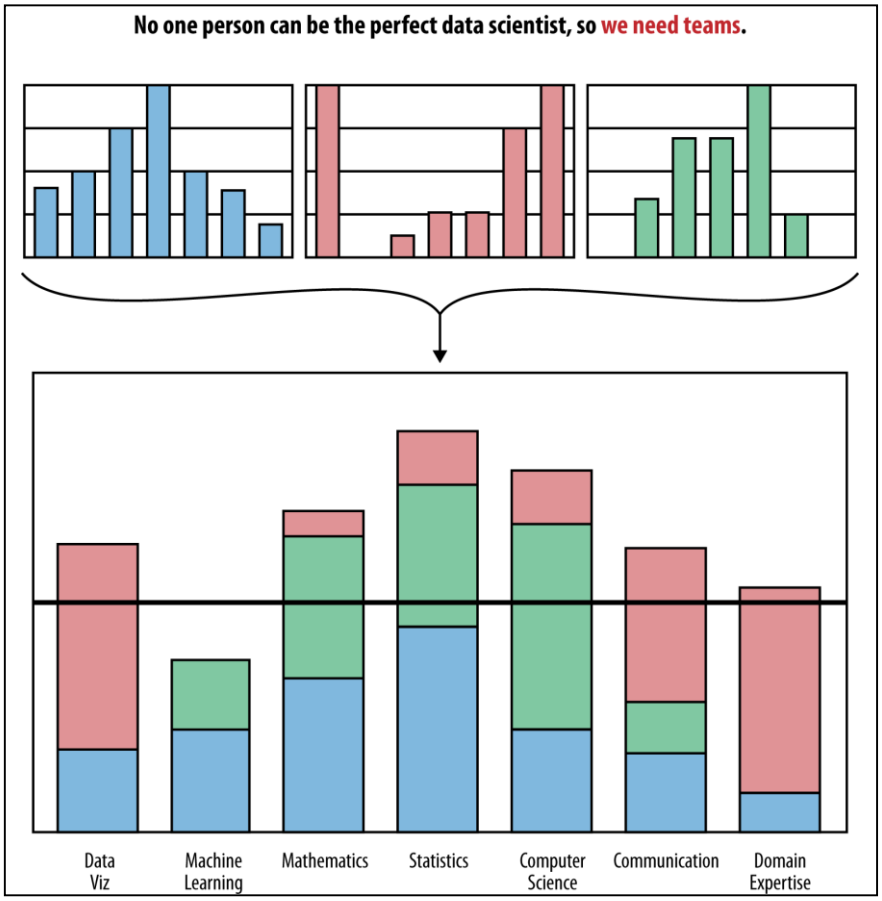


Figure 59 – Constitution d'une équipe de data science en sélectionnant divers profiles (tirée de Schutt & O'Neil, 2014)

méthodologies de modélisation des construits complexes, participer aux progrès de la mesure de l'UX. Comme toute autre science naturelle et empirique, le progrès de la psychologie a été lié de très près au développement de nouvelles méthodes permettant une mesure plus raffinée des concepts psychologiques. Les nouveaux développements techniques (mesures physiologiques, imagerie cérébrale, …) nous donnent accès aux processus psychiques de manière inimaginable pour des psychologues de la fin du 19eme siècle, le moment où ces derniers déclarèrent avoir fondé une science empirique à part entière. Les technologies informatiques et les outils de capture modernes permettent aux méthodes psychologiques classiques (observation comportementale, analyse verbale, temps de réaction, …) d'atteindre leur plein potentiel, en rendant possible le vieux rêve de mesure des individus dans leur vie de tous les jours, loin du contexte artificiel et intrusif du laboratoire. Les outils de communication modernes, comme internet ou les mobiles, permettent d'avoir des données en tous lieux, toutes heures et tout autour du globe, ce qui permet d'accéder à un nombre incommensurable de sujets potentiels d'études et correspondant à toute sa diversité réelle. La puissance de calcul actuelle permet l'utilisation d'outils statistiques puissants, permettant de manipuler des modèles élaborés et plus proches de la complexité réelle des entités étudiées. Néanmoins, il faudra prendre le temps d'intégrer tous ces nouveaux outils avec sagesse et sans brûler les étapes. La mise au point de nouvelles mesures est un processus long, et il faut se rappeler qu'il a fallu plusieurs siècles pour mettre au point un thermomètre fiable (Bryson, 2003). La mesure d'une entité encore plus insaisissable comme l'UX peut ainsi paraître d'autant plus inatteignable, mais nous pouvons nous appuyer sur ces nouveaux outils et l'expertise méthodologique passés pour mener à bien cette nouvelle entreprise.

CHAPITRE 7 : LE PROCESSUS DE CREATION D'UNE MESURE UX SOLIDE

« *You cannot manage what you cannot measure.*
You cannot measure what you cannot define.
You cannot define what you cannot understand »
John Reh

La psychologie est passée maîtresse dans la conceptualisation d'entités abstraites et leurs mesures. Son travail se base sur l'étude de divers « construits » qui correspond à « *un attribut humain que l'on postule, et dont on suppose qu'il est reflété par la performance au test* » (Cronbach & Meehl, 1955, p. 283). Chaque discipline possède les siens : la charge cognitive pour les facteurs humains, l'intelligence ou la personnalité pour la psychologie ou encore les classes sociales en sociologie (Robert & Lesage, 2011). Quand un ergonome met au point la mesure d'un construit, il doit, comme les psychométriciens, partir du qualitatif pour aller vers le quantitatif. L'ergonome est donc à la croisée des deux paradigmes, sans dogmatisme pour aucun. Il part de la conceptualisation d'un construit, puis met au point son opérationnalisation, pour enfin réaliser sa validation. De plus, dans le contexte de la mesure d'un construit abstrait comme l'UX, nous pensons qu'il convient de rajouter une dernière étape : la triangulation.

La conceptualisation : le développement de la définition conceptuelle d'un construit

« *Si j'avais le pouvoir, je commencerais par redonner leur sens aux mots* » Confucius

Dans chaque discipline, la définition d'un construit est un challenge. L'expérience utilisateur n'est pas l'exception. (Robert & Lesage, 2011). Elle est néanmoins indispensable. Pour John Reh, *« Vous ne pouvez pas gérer ce que vous ne pouvez pas mesurer, vous ne pouvez pas mesurer ce que vous ne pouvez pas définir, et vous ne pouvez pas définir ce que vous ne pouvez pas comprendre* ». C'est pour cette raison que de nombreux auteurs expliquent le retard dans la formalisation de mesures UX fiables à cause de l'ambiguïté des assises conceptuelles, elles-mêmes devant s'appuyer sur des bases théoriques solides, grâce à la modélisation (Law & Abrahão, 2014; Law et al., 2014; Law, 2011).

Définir le domaine conceptuel du construit, consiste, non pas seulement à identifier ce que le construit doit représenter ou capturer, mais également en quoi il diffère d'autres construits proches (Nunnally & Bernstein, 1994). Plus spécifiquement, il convient de spécifier la nature du construit et son thème conceptuel dans des termes non ambigus et en continuité avec la littérature existante (MacKenzie, 2003). Même si ces points sont très importants, ils sont souvent négligés et traités de manière superficielle, de nombreux praticiens pensant que nommer est équivalent à définir (MacKenzie, Podsakoff, & Podsakoff, 2011). Nous voyons ainsi fleurir de nouveaux construits sans savoir en quoi ils diffèrent de ceux qui existent déjà.

En effet, quelles différences effectives existe-t-il entre le « flow » (Cowley, Charles, Black, & Hickey, 2008; Hektner et al., 2007), l'immersion (C Jennett et al., 2008), la présence (Wijnand Ijsselsteijn & Riva, 2003; Giuseppe Riva, Waterworth, & Waterworth, 2004), l'absorption cognitive (Agarwal & Karahanna, 2000; Saadé & Bahli, 2005), la jouissance (« enjoyment » ; Vorderer, Klimmt, & Ritterfeld, 2004; Weber, Tamborini, Westcott-Baker, & Kantor, 2009), l'engagement (O'Brien & Toms, 2008; Jane Webster & Ahuja, 2006), l'implication (W. Lee, Chiu, Liu, & Chen, 2011; Tung & Deng, 2006), l'intérêt (Thiran, Bourlard, & Marques, 2010) ou l'expérience ludique ? (Mirza-Babaei & McAllister, 2010; Poels, Ijsselsteijn, & Kort, 2008). Il faut noter toutefois que la confusion entre construits est aussi vieille que la psychologie elle-même et est connue sous le nom de « *jingle–jangle fallacy* » (biais de tintement–cliquetis ; Marsh, 1994). La « *Jingle fallacy* » se rapporte à la croyance qu'un même nom d'échelle mesure forcement le même construit ; et la « *Jangle fallacy* » se rapporte à la croyance qu'un nom différent d'échelle mesure forcément un construit différent. Avec la multiplication des construits dans le domaine de l'UX, nous constatons une augmentation de ces deux biais. Durant l'enquête de Bargas-Avila & Hornbæk (2011)**,** ces derniers constatèrent une explosion conceptuelle des dimensions de l'UX, avec l'ajout de construits tels que l'enchantement (Chonchuir & McCarthy, 2008), la « *magie tangible* » (Xu, Read, Sim, McManus, & Qualter, 2009) ou encore l'esthétique d'interaction (Wakkary & Hatala, 2006), sans établissement de liens clairs avec les construits existants. En quoi l'enchantement est-il un composant crucial et distinct des autres qualités de l'UX ? Quelle est la différence entre l'enchantement et un concept, déjà plus établi, comme le flow ? Est-ce que le flow est une condition pour l'enchantement ou l'inverse ? Si de nouveaux construits UX sont proposés, ils doivent au minimum être accompagnés d'études qui clarifient leurs liens et les différences avec les construits existants, pour ne pas se retrouver avec un nombre de mots sans fin décrivant les mêmes phénomènes (Bargas-Avila & Hornbæk, 2011).

Ensuite, il s'agit de choisir le bon niveau d'abstraction et de détails pour la tâche à accomplir. En effet, il n'existe pas de délimitation immanente et naturelle d'un construit psychologique. Par exemple, dans le domaine de la psychométrie, on dénombre plusieurs centaines de tests d'habilités cognitives (Lubinski, 2006). Au fil des années, de nombreuses propositions furent faites pour les organiser. Ils furent classés au travers de 120 catégories (Guilford, 1967), sept dimensions primaires (Thurstone, 1938) ou encore d'une seule dimension dominante (Spearman, 1904). Au fur et à mesure de l'accumulation des preuves empiriques, il apparut clairement que l'organisation des habilités cognitives humaines n'était ni unitaire, ni disposée en modules spécifiques (Snow, 1986), mais hiérarchiques. L'organisation hiérarchique des capacités cognitives la plus aboutie à ce jour provient d'une synthèse de plus de 460 sets d'analyses factorielles, collectés sur la plus grande partie du siècle dernier (Carroll, 1993). Ce modèle hiérarchique en 3 strates, contient environ 60 facteurs de premier ordre, huit facteurs de deuxième ordre, et un facteur d'intelligence général à son sommet (le fameux facteur « *g* »). Cette structure hiérarchique fut corroborée par Snow et ses étudiants (Gustafsson & Snow, 1997; Marshalek, Lahman, & Snow, 1983) et la figure 60 montre un exemple de représentations graphiques des relations entre un certain nombre de tests cognitifs. Dans le domaine de l'UX un grand nombre de qualités UX ont également été collectées lors que l'enquête de Law et al. (2014). Parmi les qualités instrumentales (INQ), se retrouvent, par ordre décroissant

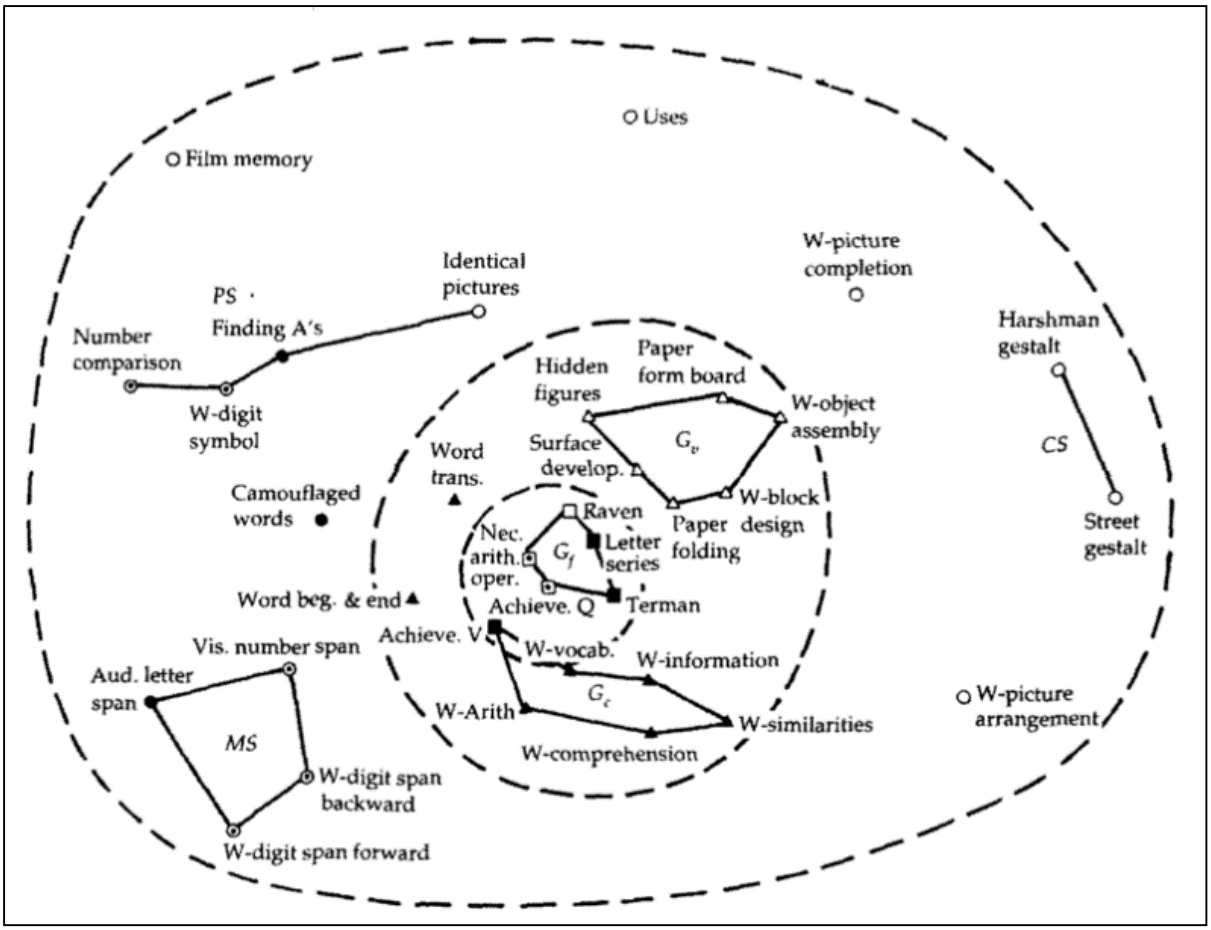


Figure 60 – Représentation concentrique du modèle hiérarchique de l'intelligence (tirée de Marshalek et al., 1983). Chaque point représente un test. Les tests sont organisés par contenus et par complexité. Les tests complexes, intermédiaires et simples sont représentés par carrés, des triangles et des cercles. Le contenu est représenté par du noir (verbal), des points (numérique) et du blanc (figuratif/spatial). Les groupes d'habilité bien connus sont indiqués par la lettre G. G_i représente l'intelligence fluide, G_c l'intelligence cristallisée, G_v l'intelligence spatiale et G_f le facteur g. Les tests les plus complexes se situent au centre de la représentation concentrique.

l'occurrence, l'utilisabilité, l'efficience, l'utilité, le temps sur la tâche, l'intuitivité, le contrôle, la clarté, le confort, le temps de réponse, l'apprenabilité, la fiabilité ou encore la vitesse. Parmi les qualités instrumentales (NIQ), il y la beauté, le challenge, la stimulation, le sens, l'attractivité, l'identification, la désirabilité ou encore la créativité. Law et al. (2014) s'interrogent sur l'accumulation de ces qualités UX ne reposant en général sur aucune théorie sous-jacente et s'apparentant plus à de la philatélie (« *Stamp-collecting* » ; p. 539). Au contraire, ces derniers pensent qu'il vaut mieux s'appuyer sur des prédicteurs plus en aval (« downstream prédictor ») et donc plus généraux et moins soumis à la contingence. Il vaut mieux ainsi se concentrer sur des construits de plus hauts niveaux, résultant d'une série de qualités UX en amont. De nombreux concepts sont concernés par cette définition. Pour les trouver, il faut aller interroger les différents modèles existants prenant en compte cette chaîne de facteurs et choisir ceux plus en aval dans le processus évaluatif de l'UX. Par exemple, dans le domaine de la psychologie des émotions, Fenouillet (2012) place le « flow » dans la dernière catégorie de son modèle intégratif de la motivation (Figure 61). La sélection de ce type de construits permet ainsi d'économiser l'étude des catégories motivationnelles précédentes (Fenouillet en dénombre 97) dont l'importance varie en fonction de la personne et du contexte d'utilisation. Des modèles similaires existent dans le domaine des IHM (Figure 62) ou dans celui plus large des medias

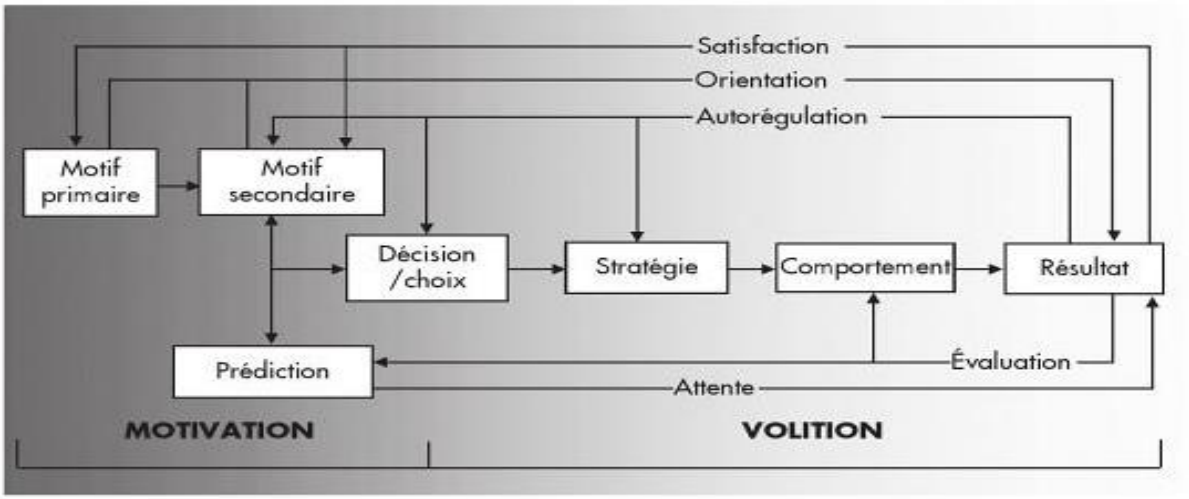


Figure 61 - Modèle intégratif de la motivation (tirée de Fenouillet, 2012)

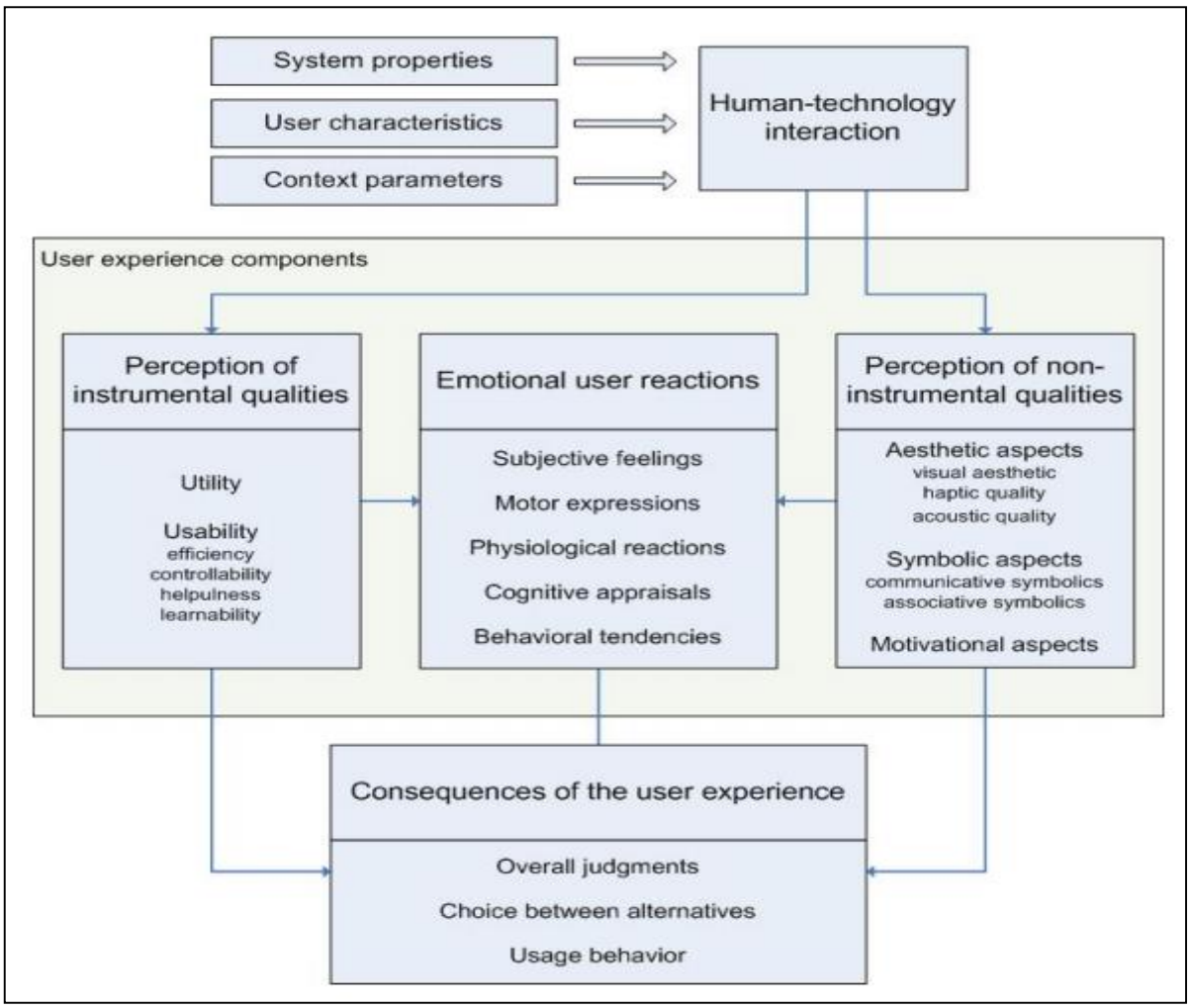


Figure 62 - Modèle de l'expérience utilisateur dans les IHM (tirée de *Mahlke, 2008)*

(Figure 63). Pour Mahlke (2008), l'interaction dépend de divers facteurs : les caractéristiques du système, l'utilisateur ainsi que le contexte d'utilisation, influencé par les facteurs sociaux, culturels et organisationnels. De cette interaction, l'utilisateur va prendre conscience des qualités instrumentales (utilité et utilisabilité) et non instrumentales du produit (aspect esthétique, symbolique et motivationnel), ce qui va déclencher des réactions affectives chez

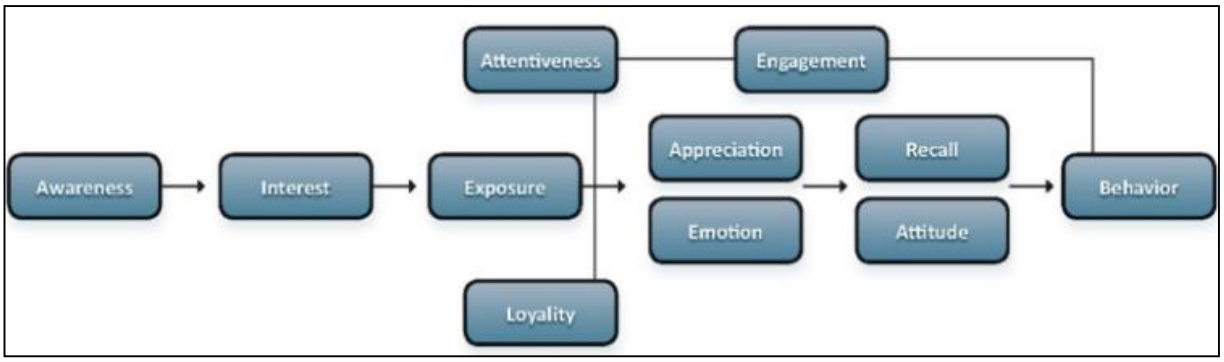


Figure 63- Modèle des dimensions de l'audience (Napoli, 2010)

l'utilisateur et influer sur son comportement. Pour Napoli (2010) l'expérience médiatique commence par la prise de connaissance du média, qui peut alors conduire à l'intérêt et l'exposition. S'ensuit l'engagement qui comprend l'appréciation et la réponse émotionnelle. Ceux-ci influencent à leur tour le rappel et l'attitude qui conduisent à des changements de comportement. Dans ces modèles d'expérience utilisateur, se retrouvent couramment en construit « *de fin de chaine* » les notions de flow, d'immersion, d'absorption ou d'engagement. Ce sont des construits particulièrement intéressants pour la mise au point d'une mesure globale de l'UX.

L'opérationnalisation : le choix des mesures et leurs combinaisons

Il existe de nombreuses façons d'opérationnaliser la mesure d'un construit telle que l'expérience utilisateur. La plus classique se base sur l'interrogation verbale d'individus à partir d'un questionnaire. Il est également possible d'inférer l'état d'un sujet à partir de son état physiologique ou de ses traces comportementales. Enfin, il est possible de combiner ces approches pour pallier les faiblesses de chacune.

La méthode verbale : le recueil de trace écrite ou de déclaration auto-rapportée

La méthode la plus simple pour recueillir des données subjectives est simplement d'interroger la personne ciblée. Les méthodes d'auto-évaluation offrent des avantages évidents par rapport aux autres techniques d'évaluation. Ces méthodes sont simples, rapides, peu coûteuses, flexibles, et fournissent souvent des informations qui seraient difficiles ou impossibles à obtenir autrement (Lucas & Baird, 2006). Les gens peuvent généralement auto-rapporter leur expérience d'interaction dans le cadre d'un entretien ou en remplissant un questionnaire. Cependant, ces techniques comportent également des pièges qui requièrent une prudence particulière lors de leurs utilisations. De nombreux chercheurs déplorent le manque de correspondance entre ce que les gens disent faire et ce qu'ils font réellement. De même, il y a un manque d'analogie entre ce que les personnes disent ressentir et ce que leur activité psychophysiologique indique qu'elles ressentent probablement. Par conséquent, ce que les gens disent à propos de leur état subjectif est parfois différent de ce que les expressions comportementales et physiologiques suggèrent. Les raisons sont multiples et ces différences

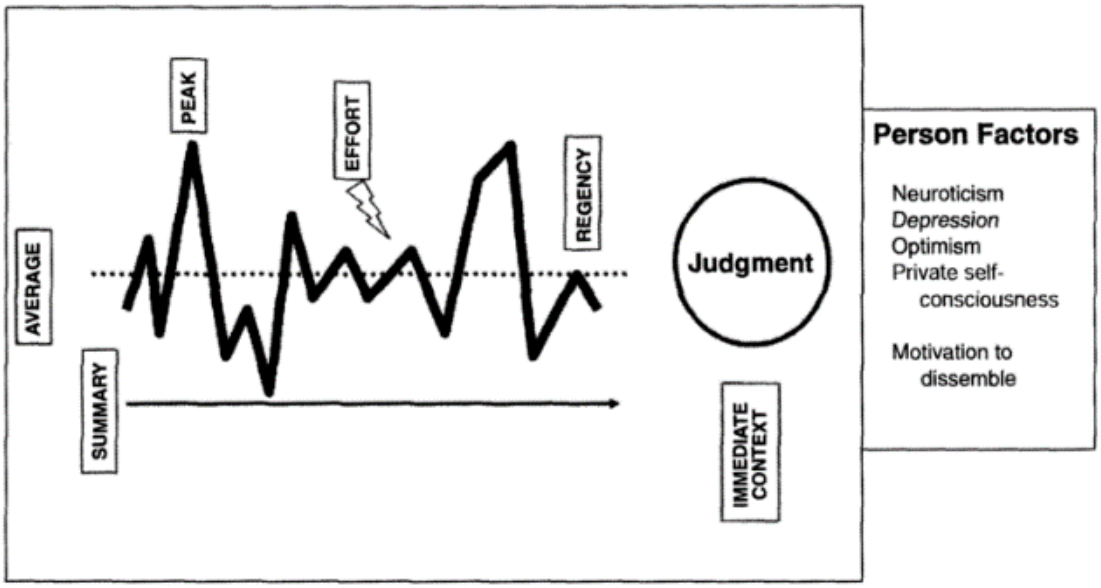


Figure 64 - Facteurs pouvant un jugement de rappel (Stone & Litcher-kelly, 2006)

sont d'autant plus importantes quand on demande à un sujet d'évaluer un état passé. La figure 64 expose un certain nombre des facteurs pouvant influencer ce jugement (Stone & Litcher-kelly, 2006). En raison de ces divergences, les chercheurs essayent de plus en plus de recueillir d'autres types de mesures en même temps.

Une des interrogations méthodologiques fortes concerne donc le moment où l'on va solliciter l'utilisateur pour qu'il nous fasse part de son état subjectif. La méthode la plus classique est de demander à l'utilisateur de remplir un questionnaire après l'expérience d'interaction. Si cette méthode est présentée comme la plus fiable, car le questionnaire, souvent conséquent, permet une évaluation du construit très précise, elle a pour inconvénient de se situer après l'interaction et donc soumise aux biais de jugement présenté par Stone & Litcher-kelly (2006). D'autres échelles ont été mises au point, plus courtes (Finstad, 2010, 2013), voire composées d'un seul item (Christophersen & Konradt, 2011). Ces mesures possèdent une fiabilité et une validité correctes, tout en permettant, par leur rapidité de mise en œuvre, une utilisation multiple lors de la situation d'interaction, couplée à des coupures moins longues entre la tâche d'évaluation et la situation d'étude. D'autres techniques mises au point pour limiter davantage la coupure entre les deux tâches, a été d'intégrer directement dans l'interface un bouton de feedback affectif (Broekens & Brinkman, 2013) ou d'utiliser un dispositif physique dédié exclusivement à l'estimation de son propre niveau de présence (IJsselsteijn, de Ridder, Hamberg, Bouwhuis, & Freeman, 1998). Si ces techniques ont l'avantage de mesurer un construit de manière continue, elles ont comme inconvénient de distraire les sujets de la tâche et de demander continuellement des actions supplémentaires de leur part (Chung & Gardner, 2012).

D'autres mesures de l'UX peuvent se faire sans que l'utilisateur en ait conscience à partir de l'analyse de ses traces verbales, via des techniques de datamining et de crowd-sourcing (Alkhattabi, Neagu, & Cullen, 2011; Andresen, 2009; Tang & Boring, 2012), puis complétées par une analyse de contenus (Donnelly & Gardner, 2011; Mehl, 2006; Traphagan et al., 2010), de sentiments (De Choudhury, Gamon, & Counts, 2012; Grassi, Cambria, Hussain, & Piazza, 2011; Nahin, Alam, Mahmud, & Hasan, 2014; Neviarouskaya, Prendinger, & Ishizuka, 2010;

Pan, Ni, Sun, Yang, & Chen, 2010; L. Zhang & Yap, 2012), de charge cognitive (Khawaja, Chen, & Marcus, 2014) ou encore d'engagement (D'Mello & Graesser, 2010). Certaines initiatives tentent même d'effectuer cette analyse en temps réel, via l'utilisation de « backchannel », un espace d'expression créé à côté de l'activité et permettant à ceux qui le suivent de réagir (Huron, Isenberg, & Fekete, 2013). Un avantage majeur de l'utilisation de ces techniques est leur caractère non réactif et écologique (Fritsche & Linneweber, 2006), c'est-à-dire que les données recueillies ne sont pas biaisées comme peuvent l'être celles qui sont obtenues dans un contexte artificiel et provoqué (comme dans un test en laboratoire). De plus, les traces, si elles sont datés et nombreuses, peuvent nous donner des indications fines sur l'activité tout au long de celle-ci. Néanmoins, ces données peuvent être éparses et bruitées car elles ne sont ni provoquées ni contrôlées ; et peuvent être également non représentatives car l'échantillonnage n'est généralement pas contrôlé. Enfin, certaines données peuvent être tout simplement fausses, bien que des techniques sont actuellement à l'épreuve pour détecter les commentaires fallacieux fabriqués de toutes pièces (Mukherjee, Liu, & Glance, 2012).

Les mesures physiologiques

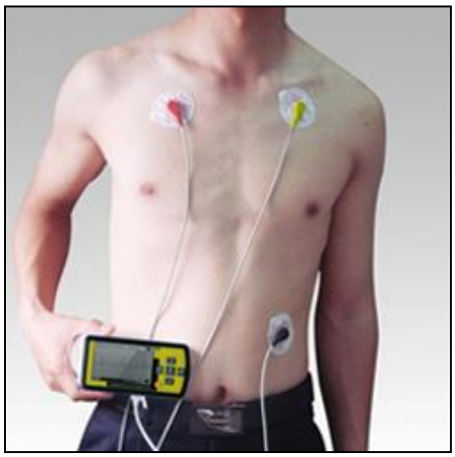


Figure 65 – Electrocardiogramme (tirée de tmsmedicaltechnologies.com)

Une autre façon d'enrichir la mesure de l'UX est de mettre à profit la captation de certains signaux de notre corps. En effet, il est possible d'inférer certains états mentaux à partir de signaux physiologiques. La discipline qui étudie le changement de l'activité physiologique en fonction de certains déclencheurs psychologiques s'appelle la psychophysiologie, qui est une branche de la psychologie (Cacioppo, Tassinary, & Berntson, 2007). Lorsque les gens et les animaux se préparent à s'engager dans diverses activités, le système nerveux et endocrinien commence à fabriquer et à libérer diverses substances chimiques qui soutiennent l'activité biologique lors d'états immersifs et affectifs. Cela entraîne divers bouleversements physiologiques, tels que la modification de la fréquence cardiaque ou la dilatation de la pupille. C'est en exploitant ces modifications que l'on peut inférer certains états psychologiques.

L'une des familles d'indicateurs physiologiques la plus exploitée concerne le système cardio-vasculaire, qui a pour rôle de réguler le flux sanguin dans tout le corps. Ses variations peuvent refléter des changements affectifs et cognitifs. Les mesures de l'activité cardio-vasculaire les plus courantes sont le rythme cardiaque (HR), la variabilité du rythme cardiaque (HRV) ou encore la pulsation du volume sanguin. Le rythme cardiaque et la variabilité du rythme cardiaque peuvent être collectés à partir d'un électrocardiogramme (ECG ; Figure 65). Pour capturer le signal cardiaque, deux électrodes éloignées l'une de l'autre sont placées sur le sujet, en général sur la poitrine et sur l'abdomen (Stern, Ray, & Quigley, 2001). La pulsation du rythme cardiaque est capturée grâce à un photopléthysmographe (PPG ; Figure 66). Cet appareil

Figure 66 – Photopléthysmographe (tirée de nonin.com)

repose sur une technique de mesure optique non invasive, qui détermine le rythme cardiaque en repérant les fines variations de la lumière reflétée par la peau causées par les fluctuations du volume sanguin. Le rythme cardiaque (HV), tout comme la pulsation du rythme cardiaque (BVP), sont des indicateurs classiques de l'état émotionnel et de l'activité mentale. Ils augmentent généralement pendant l'excitation, la concentration et la présentation d'un stimulus sensoriel intenses ; et diminuent lors de la relaxation, l'observation visuelle et auditive attentive, et le traitement d'un stimulus plaisant (Frijda, 1986). Avec le contrôle supplémentaire de la température des doigts, il serait également possible de différencier les émotions positives et négatives (Papillo & Shapiro, 1990). La variabilité du rythme cardiaque (HRV) est un indicateur riche de l'activité mentale. Sous une situation de stress, l'HRV baisse et quand le sujet est reposé, elle augmente. De plus, sous un effort mental, l'HRV diminue, mais si l'effort dépasse les capacités de traitement du sujet, ce même indicateur augmente (Rowe, Sibert, & Irwin, 1998). Ce sont des mesures qui ont été très utilisées pour mesurer l'engagement émotionnel (Niklas Ravaja, Saari, Salminen, Laarni, & Kallinen, 2006; Sammler, Grigutsch, Fritz, & Koelsch, 2007; Siefert et al., 2009; R. D. Ward & Marsden, 2003), ludique (Cui & Rau, 2012; Mandryk, Inkpen, & Calvert, 2006; Money & Agius, 2009) et « *optimal* », par le concept de « *flow* » (Keller, Bless, Blomann, & Kleinböhl, 2011; Peifer, Schulz, Schächinger, Baumann, & Antoni, 2014; Ulle, 2010). L'activité plasmique, par la variation du contenu des flux sanguins, peut également nous renseigner sur l'état mental de l'utilisateur, tels que le taux de cortisol pour le « *flow* » (Keller et al., 2011; Peifer et al., 2014), ou des catécholamines et des leucocytes pour l'engagement mental (D. E. Brown, James, Nordloh, & Jones, 2003; Shelton-Rayner, Mian, Chandler, Robertson, & Macdonald, 2012). Ces mesures varient en intrusivité, le recueil de fluide étant la méthode la plus extrême. Les techniques les moins contraignantes telles que le PPG pouvant se matérialiser sous la forme d'un clips à l'oreille, ont comme désavantage de présenter la plus grande vulnérabilité aux artéfacts causés par des mouvements parasites.

Le comportement de l'œil, tel que la taille de la pupille, la fréquence des clins d'œil et les patterns des mouvements oculaires, peuvent également nous renseigner sur certains de nos états psychologiques. Ces données sont captées à partir d'un eye-tracker (Figure 67), dont les plus courants utilisent un système d'enregistrement infrarouge pour capter le mouvement des yeux, la réflexion cornéenne et la taille de la pupille. À partir de ces informations, l'eye-tracker suit le mouvement des yeux et infèrent la position du regard sur l'écran à partir d'une calibration initiale. La dilatation pupillaire, qui agit sur sa taille, est un bon indicateur de l'activité mentale, de l'intérêt et de l'intensité des émotions. Ainsi, Hess & Polt (1960) ont montré une dilatation d'environ 20%, par rapport au niveau de base, du diamètre pupillaire chez des personnes visionnant des images intéressantes pour eux. Plus cette stimulation est intense, plus la réponse pupillaire

Figure 67 – Eyetracker fixe et mobile (tirée de http://certesens.univ-tours.fr/)

est forte (Just & Carpenter, 1993; Kahneman, 1973). De plus, la dilatation pupillaire se maintient tant que l'intérêt est conservé (White & Maltzman, 1978). C'est pourquoi cette mesure a été souvent utilisée pour évaluer l'engagement cognitif (Ahlstrom & Friedman-Berg, 2006; Chen & Epps, 2014; Piquado, Isaacowitz, & Wingfield, 2010; Wang, Duffy, & Du, 2007) et affectif (Bradley, Miccoli, Escrig, & Lang, 2008; Gao, Barreto, & Adjouadi, 2009; Partalaa & Surakka, 2003), la nouveauté (Naber, Frässle, Rutishauser, & Einhäuser, 2013) ou encore la vigilance (Mcintire et al., 2011). Cependant, la pupillométrie est difficile à mettre en place dans un contexte d'éclairage naturel, car celle-ci est perturbée par les changements de luminosité qui sont difficiles à anticiper et à maitriser en dehors des situations de total contrôle (Ganglbauer, Schrammel, Deutsch, & Tscheligi, 2011). Un autre indicateur oculaire de l'engagement est la fréquence de clignement de l'œil, qui est associée régulièrement à l'augmentation de la charge cognitive (Ahlstrom & Friedman-Berg, 2006; S. Chen & Epps, 2014; Nourbakhsh, Wang, & Chen, 2013; Siegle, Ichikawa, & Steinhauer, 2008; Wang et al., 2007) et affective (Adam, Mallan, & Lipp, 2009; Cui & Rau, 2012; Kapoor, Burleson, & Picard, 2007). Enfin, la direction du regard en lui-même et ses variations sont d'excellents indicateurs de l'intérêt. De ce fait, Les fixations et les mouvements oculaires ont été utilisés pour mesurer la pertinence d'un objet (Buscher, Dengel, & van Elst, 2008) et son intérêt (Ajanki, Hardoon, Kaski, Puolamäki, & Shawe-Taylor, 2009), ainsi que pour déterminer un niveau d'attention (Frischen, Bayliss, & Tipper, 2007), d'engagement (Nakano & Ishii, 2010) ou d'immersion d'un utilisateur (Charlene Jennett, Cox, & Cairns, 2009; Pivec & Pivec, 2009). Ces mesures ont un grand potentiel, même si elles demandent encore l'acquisition de dispositifs coûteux et d'un calibrage avant chaque utilisation.

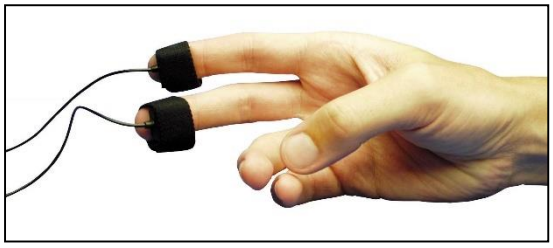


Figure 68 – Capteurs électrodermaux (tirée de biopac)

L'activité électrodermale est une famille d'indicateurs psychophysiologiques intéressants pour la capture de l'UX. La plus utilisée est la réponse galvanique (GSR), qui correspond aux changements électriques à la surface de la peau. Tous les stimuli nouveaux, émotionnels, menaçant et attirant l'attention, activent des glandes sous-cutanées qui créent une micro-sudation à la surface de la peau, ce qui change sa conductivité électrique. C'est ce changement de conductance de la peau qui est mesuré grâce à des électrodes (Figure 68) placées en général sur les doigts de la main de l'utilisateur qui ne participe pas à l'activité, afin de ne pas contraindre ses mouvements et de ne pas parasiter le signal. La réponse galvanique de la peau est corrélée linéairement avec une excitation à la fois émotionnelle (Grewe, Nagel, Kopiez, & Altenmu, 2007; Kapoor et al., 2007; Lang, Greenwald, Bradley, & Hamm, 1993; Peter & Herbon, 2006; Ward & Marsden, 2003) et cognitive (Boucsein, 1992; Haapalainen, Kim, Forlizzi, & Dey, 2010; Miyake et al., 2009), ce qui en fait un bon indicateur de l'engagement (Latulipe, Carroll, & Lottridge, 2011; Sundar, Xu, Bellur, Oh, & Jia, 2010) Cet indicateur est couramment utilisé dans l'évaluation du rythme immersif dans les jeux vidéo (Ivory & Magee, 2009; Kivikangas et al., 2010; Mandryk et al., 2006; Mirza-Babaei et al., 2012; Ravaja et al., 2006) . Il permet ainsi de cerner les conditions d'immersion dans un environnement numérique et les moments d'excitation et d'ennui qui l'accompagnent. Il permet l'enregistrement de

l'évolution des états d'investissement du joueur avec une granularité très fine, sans demander d'interruption de l'activité en cours et de qualités introspectives particulières de la part des utilisateurs (Léger, Fouquereau, Levillain, & Tijus, 2009). Cependant, une difficulté inhérente au recueil de ce type de mesures est son faible niveau de discrimination à l'égard des qualités phénoménales des émotions éprouvées par un individu, comme de distinguer entre un épisode de frustration ou d'excitation.

L'activité squelettique et musculaire, comme avec les expressions faciales et posturales, peuvent également nous renseigner sur l'expérience utilisateur lors d'une interaction. Ce sont des indicateurs qui sont classés parfois en tant qu'indicateurs physiologiques et parfois en tant qu'indicateurs comportementaux. Les expressions faciales sont des indices naturels de l'état émotionnel d'une personne. Elles sont capturées généralement soit par électromyogramme (EMG), soit à partir d'une analyse vidéo. L'EMG faciale (Figure 69) est fréquemment utilisée pour indiquer la valence positive ou négative d'un état émotionnel (P.J. Lang et al., 1993). L'activité enregistrée se base sur trois groupes de muscles du visage : le zygomaticus major (ZM), le corrugator supercilii (CS) et l'orbicularis oculi (OO). L'activation du zygomaticus major (ZM), le muscle du sourire, et du corrugator supercilii (CS), le muscle de froncement de sourcils ont été associés à des émotions positives et négatives , respectivement (Lang et al., 1993; Witvliet & Vrana, 1995). Lors de la présentation de stimulus déplaisants, les muscles CS montrent une activité accrue (Ravaja, 2004). L'activation de la région externe de l'orbicularis oculi (OO) est associée au sourire vrai (Ekman, Davidson, & Friesen, 1990) et au plaisir intense (Jäncke, 1994; Ravaja, 2004). L'EMG a été utilisé dans de nombreuses études pour mesurer la valence émotionnelle lors de la présentation de stimulus auditifs (Grewe et al., 2007; Kallinen & Ravaja, 2005; Nacke, Grimshaw, & Lindley, 2010), vidéos (Partala, Surakka, & Vanhala, 2006), interactifs (Hazlett & Benedek, 2007; Tuch, Bargas-Avila, Opwis, & Wilhelm, 2009) ou ludiques (Kivikangas et al., 2010; Mandryk et al., 2006; Nacke et al., 2010; Ravaja, 2008). L'avantage de l'EMG faciale est sa grande précision qui lui permet de détecter les mouvements musculaires qui sont trop petits ou rapides pour être décelés visuellement. Néanmoins, l'utilisation de cette technique est qu'elle implique l'utilisation d'électrodes fixés sur le visage, ce qui est dérangeant, très invasif , et nécessite des compétences de manipulation et d'analyse spécialisées (Ward, 2004). Depuis les années 90, le développement fulgurant de la vision numérique (Essa, 1999; Yang & Ahuja, 2001), de la vidéo numérique et de la puissance de calcul des ordinateurs a permis l'essor d'une nouvelle technique de recueil des expressions faciales à partir de systèmes fondés sur l'analyse de données vidéo (Mishra et al., 2015; Pantic & Rothkrantz, 2000). Cette technique, plus flexible et moins intrusive a été utilisée pour détecter les émotions à partir des expression faciales (Bailenson et al., 2008; Kapoor et al., 2007). Grâce à des indices issus du langage corporel, elle a été également capable de mesurer d'autres états mentaux, tels que l'intérêt (Mota & Picard, 2003), l'attention (Dirican & Göktürk, 2012; Lee et al., 2006), l'engagement (Asteriadis, Karpouzis,

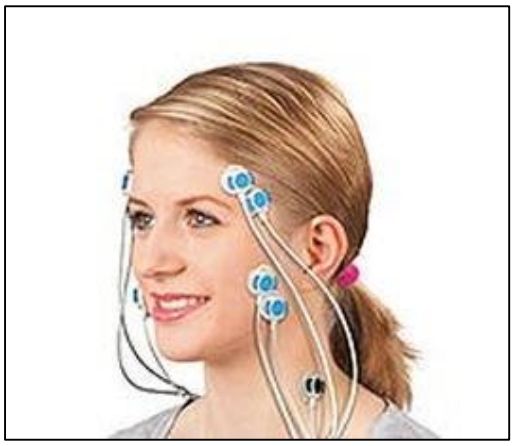


Figure 69 – EMG faciale (tirée de medicalexpo.fr)

& Kollias, 2009; D'Mello & Graesser, 2010) ou encore le flow (Ulle, 2010). De plus, des capteurs embarqués dans une chaise de bureau ont permis la reconnaissance automatique de l'état d'attention (Mutlu, Krause, Forlizzi, Guestrin, & Hodgins, 2007) et de frustration (Kapoor et al., 2007), à partir de l'analyse de la posture assise de l'utilisateur. La reconnaissance des états internes par l'analyse vidéo des expressions faciales et posturales a pour limite actuelle de demander des conditions d'éclairage spécifiques et une qualité vidéo suffisante, ce qui limite la qualité des données recueillies dans des environnements de capture non contrôlés.

Enfin, la méthode la plus directe pour mesurer l'état mental d'une personne consiste à observer l'activité de diverses régions cérébrales, telles que le cortex cérébral ou le système limbique. Les dispositifs les plus utilisés pour cela sont le scanner à résonance magnétique (fMRI ; Figure 70) et l'électroencéphalogramme (EEG ; Figure 71). L'imagerie par résonance magnétique permise par le fMRI est une technique extrêmement puissante et précise qui permet de mesurer la structure, la fonction, la connectivité et la composition chimique de n'importe quelle partie de notre corps. La localisation des zones cérébrales activées est basée sur l'effet BOLD (« *Blood Oxygen Level Dependant* »), lié à l'aimantation de l'hémoglobine contenue dans les globules rouges du sang. L'expérience plaisante ressentie lors de l'accomplissement d'une tâche est liée au système limbique, et plus particulièrement au striatum, qui fait partie du système de récompense (Baldo & Kelley, 2007; Frackowiak et al., 2003; Lang & Bradley, 2010). Par exemple, Baldo et Kelley (2007) ont montré un lien entre la libération de neuromédiateurs au niveau du stratium et l'expérience de récompense ressentie lors d'un comportement de consommation. Les réseaux d'alerte et de direction de l'attention (Fan, McCandliss, Fossella, Flombaum, & Posner, 2005) sont également très intéressants pour la mesure de l'UX, car le maintien de l'attention renseigne sur l'intérêt de l'utilisateur. Ces réseaux se situent dans les régions du lobe pariétal inférieur et supérieur, du lobe frontal des activations rétinotopiques (FEF) et du colliculus inférieur (Fan, McCandliss, Sommer, Raz, & Posner, 2002). Une hypothèse intéressante à propos de l'état de flow est qu'il correspond neurologiquement en une synchronisation du réseau attentionnel et de récompense. De manière complémentaire, une étude montre que le sentiment d'effacement de soi ressenti lors de cet état d'immersion peut être détecté par l'observation de la baisse d'activité dans le cortex préfrontal médial (Ulrich, Keller, Hoenig, Waller, & Grön, 2014), qui a pour rôle de réguler la conscience de soi (D'Argembeau et al., 2007). L'imagerie cérébrale a également été utilisée pour étudier d'autres construits UX, comme le sentiment de présence (Bouchard et al., 2009) ou encore l'expérience

Figure 70 – Scanner MRI (tirée de physicscentral.com)

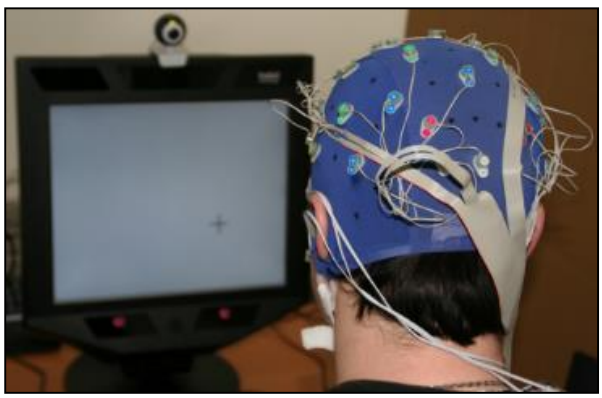


Figure 71 – Electroencephalogramme (tirée de *Nacke, 2009)*

esthétique (Reimann, Zaichkowsky, Neuhaus, Bender, & Weber, 2010). Néanmoins, même si cette méthode permet d'étudier finement les bases biologiques de certains construits abstraits, ses caractéristiques invasives, son coût de fonctionnement et la déformation de l'activité observée par les contraintes physiques de l'appareil, rend cette technique peu opérante pour la mise au point d'une mesure pratique de l'UX. De ce fait, beaucoup de chercheurs, pour des raisons de coût ou en fonction de la nature de la tâche préfèrent utiliser l'EEG pour mesurer l'activité cérébrale. C'est une méthode qui mesure l'activité électrique du cerveau par des électrodes placées sur le cuir chevelu. Les oscillations électriques sont ensuite analysées à partir de quatre bandes de fréquences, appelées bande delta (1–4 Hz), bande thêta (4–7 Hz), bande alpha (8–12 Hz) et bande beta (13–30 Hz). En général, l'activité alpha EEG est considérée comme étant liée aux demandes attentionnelles alors que l'activité bêta est liée au processus émotionnel et cognitif (Ray & Cole, 1985). Les recherches indiquent que l'augmentation de l'activité bêta est associée à un niveau plus élevé de vigilance (Freeman, Mikulka, Prinzel, & Scerbo, 1999), voire, à un niveau très élevé, de détresse (Woodruff, Daut, Brower, & Bragg, 2011). L'activité de la bande alpha est associée à un état d'éveil relaxé, de bien-être et d'absence d'anxiété (Van Boxtel et al., 2012). L'activité de la bande thêta ressemble aux « *rêves éveillés* » (« *Daydreaming* »), à la créativité, l'intuition, aux réminiscences, sensations et émotions (Aftanas & Golocheikine, 2001). Enfin, l'activité de la bande delta est importante durant le sommeil profond et peut-être associée à des processus mentaux inconscients, comme la transe (Cacioppo, Tassinary, & Berntson, 2007). Une activité dans la bande alpha importante, couplée à une activité dans la bande beta moindre, semble être un bon indicateur de l'absorption cognitive (Léger, Davis, Cronan, & Perret, 2014). C'est pourquoi l'EEG est utilisé couramment pour mesurer l'engagement (Leslie, Ojeda, & Makeig, 2014; Nacke, 2009) ou l'attention (Rebolledo-mendez, Dunwell, & Martínez-mirón, 2009; Szafir & Mutlu, 2012). Malgré des efforts récents dans le but de minimiser l'encombrement, en enchâssant par exemple des électrodes sans fil dans des lunettes de vue (Salvucci, 1999), l'intrusivité, le coût et la difficulté d'interprétation des données recueillies freine encore l'utilisation de cette méthode.

En conclusion, les mesures physiologiques ont l'avantage de ne pas partager certains biais propres aux évaluations subjectives auto-rapportées, telles que provoquées par la mémoire, les interprétations cognitives ou la désirabilité sociale (Ravaja, 2004). Mieux, elles donnent accès à des réponses émotionnelles et cognitives subtiles, parfois non accessibles à la conscience. De plus, elles permettent une capture de données fines, en temps réel et sans (trop) perturber l'activité de l'utilisateur. Néanmoins pour des raisons pratiques et méthodologiques, les outils de mesure actuels sont encore considérés comme trop intrusifs et encombrants. Leur utilisation ne permet pas d'observer l'utilisateur dans des conditions écologiques, en partie à cause du câblage, du coût temporel de la calibration des outils et de la pose de senseurs intrusifs, vécue comme une grande situation de stress pour l'utilisateur. On peut noter toutefois des progrès rapides dans ce domaine. En effet, pratiquement chacune des méthodes présentées a vu naître des dispositifs bien moins encombrants et plus simples d'utilisation, bien que cela se fasse en partie par une réduction de la précision de l'outil : des gants (Figure 72 ; Picard & Scheirer, 2001), montres (Broadwater, Haynes, & Mitry, 1982) et sièges (Figure 73; Aley et al., 2005; Anttonen, Surakka, & Koivuluoma, 2009; Schumm et al., 2010) mesurant la conductivité de la peau et le rythme cardiaque ; des Kinects (Burba, Bolas, Krum, & Suma, 2012) pour mesurer

Figure 72 – Gant « Galvactivator » , mesurant la conductivité de la peau (Picard & Scheirer, 2001).

Figure 73 – La chaise eMFI, mesurant le rythme cardiaque (Anttonen et al., 2009)

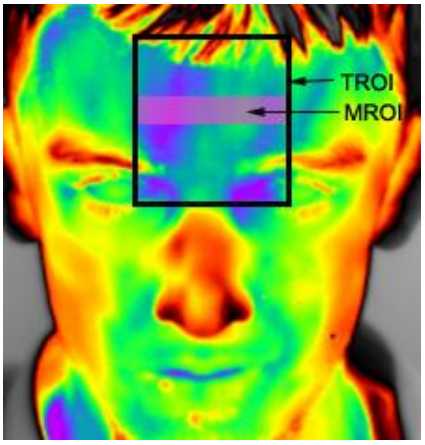


Figure 74 – image obtenu par la camera StressCam, mesurant la concentration sanguine (Yun, et al., 2009)

la respiration et le langage corporel ; des radars micro-ondes pour mesurer à distance le rythme cardiaque et la respiration (Suzuki, Matsui, Kagawa, Asao, & Kotani, 2013); des caméras thermiques fixant le front pour mesurer la charge mentale (Figure 74; Shastri, Pavlidis, & Wesley, 2009; Wang et al., 2007), la distraction (Wesley, Shastri, & Pavlidis, 2010) ou la frustration (Yun, Shastri, Pavlidis, & Deng, 2009); des EEG low-cost (Knoll, Wang, Chen, & Xu, 2011) ou à electrode réduit et sec (Rebolledo-mendez et al., 2009); des systèmes de suivi oculaires sans calibration (Kohlbecher et al., 2008; Model & Eizenman, 2012) ou à partir de webcams classiques (Johansen, San Agustin, Skovsgaard, Hansen, & Tall, 2011; Sesma, Villanueva, & Cabeza, 2012). Un autre problème potentiel avec les mesures physiologiques est qu'il est difficile d'inférer un seul état psychologique d'une seule mesure. En effet, le rythme cardiaque peut augmenter lors d'une excitation émotionnelle et baisser lors d'un engagement attentionnel. De ce fait, une augmentation du rythme cardiaque pourrait très bien être interprétée par un déficit attentionnel et sa baisse par un manque d'excitation émotionnelle. L'inférence d'un état psychologique à l'aide d'une seule variable physiologique étant difficile, un travail de triangulation des mesures entre-elles sera à faire. Dans cette approche prometteuse, Picard, Vyza et Healey (2001) ont réussi la reconnaissance de huit émotions avec une précision de 81%, en combinant les mesures de la respiration, de la pression sanguine, de la conductance de la peau et des expressions faciales. Enfin, un travail sur les mesures et les corrélations avec les entités qu'elles sont censées mesurer est à faire. Pour Fairclough (2009) ces variables physiologiques devront posséder trois grandes qualités :

- **DIAGNOSTICITE** : habilité d'une variable physiologique à cibler précisément un état psychologique sans être affecté par d'autres
- **SENSIBILITE** : habilité d'une variable physiologique à détecter rapidement et précisément un changement psychologique
- **FIABILITE** : consistance d'une inférence psycho-physiologique entre les différents individus et environnements

Les mesures comportementales

Ce genre de mesure était prépondérant dans le paradigme de l'utilisabilité, il l'est beaucoup moins dans le cadre de la mesure de l'UX. Or, de nombreux aspects du comportement peuvent exprimer l'engagement et l'intérêt d'une personne dans une activité :

- **L'EFFORT :** l'étendu de l'effort mis en avant tout en essayant d'accomplir une tâche ;
- **LA LATENCE :** le temps mis par une personne pour répondre suite à l'exposition à un stimulus ;
- **La PERSISTANCE :** le temps entre le déclenchement de la réponse et sa cessation ;
- **LE CHOIX :** lorsqu'il est présenté avec plusieurs pistes d'actions montrant une préférence pour une piste plutôt qu'une autre ;
- **LA PROBABILITE DE REPONSE :** étant donné les différentes opportunités qu'un comportement se produise, le nombre (ou pourcentage) d'occasions pour lesquelles une réponse particulière dirigée vers un but se produit ;

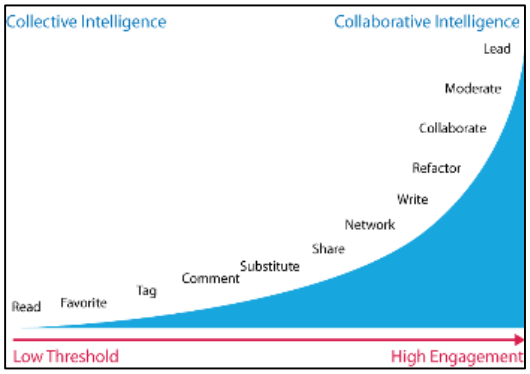


Figure 75 – Loi de la participation (power law) de Ross Mayfield (2006)

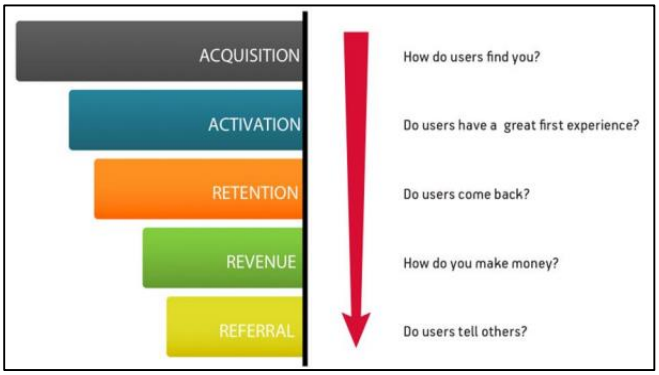


Figure 76 – Modèle Pirate de McClure (2007)

Ce type de mesure n'a jamais été vraiment exploitée, car les évaluations classiques des systèmes interactifs mettaient les utilisateurs en situation artificielle, en provoquant leurs comportements et en leurs demandant de réaliser certaines tâches. Le comportement ne reflétant pas la motivation naturelle des sujets, ce genre de mesure aurait été faussé. Or, avec l'explosion du recueil de données sur Internet, il est possible de capturer les comportements utilisateur sans que nos dispositifs de mesure ne les parasitent. Ainsi, de nombreuses métriques se basant sur le comportement utilisateur sur Internet ont été développées pour mesurer l'engagement des utilisateurs. La base de la réflexion est que plus un internaute sera engagé, plus il sera enclin à participer (Figure 75). Ainsi, plus l'utilisateur aimera l'expérience proposée par le service, plus il y passera de temps, explorera de liens, et participera en commentant ou rajoutant du contenu. Les métriques « Pirates » de Dave McClure (2007) figurent parmi les plus emblématiques d'entre elles (Olsen, 2015). Les premières lettres de chacune (Acquisition, Activation, Rétention, Revenu, Orientation) forment le mot « Aarrr », qui sonne à haute voix comme le cri de guerre d'un pirate, d'où son nom. En reprenant un entonnoir de vente traditionnelle (Figure76), ce modèle permet de mesurer à chaque étape le niveau d'engagement de plus en plus important, en se basant sur le comportement des utilisateurs lors des différents stades :

- **ACQUISITION :** plus de deux pages, deux clics, plus de dix secondes
- **ACTIVATION :** longue visite (X pages, resté Y secondes, Z clics), enregistrement de compte
- **RETENTION :** visite répétés (3+ visites par mois)

- **REFERENCE :** recommande le service à des amis
- **REVENU :** abonnement, achats

Ce genre de modèle permet de piloter la croissance d'un service numérique lors de son déploiement. Il est couramment utilisé par les growth hackers, qui l'utilisent pour voir l'impact d'une modification sur l'engagement des utilisateurs. La limite de cette approche est qu'elle rend difficile la comparaison avec la concurrence, car les données ne sont généralement pas accessibles (on ne dispose que des données de son service) et sont peu transposables entre les sites, car la structure unique de chacun d'entre eux influence spécifiquement leurs expressions. D'autres mesures se basent uniquement sur les données des réseaux sociaux, tels que les métriques d'engagement sociales de Kaushik[76], ce qui a pour avantage de permettre la comparaison entre services concurrents. Elles sont au nombre de quatre : le taux de conversation (nombre de conversations par post ; nombre de commentaires sur Facebook, Google+, LinkedIn ; nombre de réponses sur tweeter), le taux d'amplification (nombre de repartage ou de retweet par post), le taux d'applaudissement (retweet, j'aime, +1, …), et la valeur économique (la somme des revenus court terme, long terme, et d'économies d'argent). La plupart des géants du net ont mis au point leurs propres systèmes de suivi d'audience, tels que Google analytics[77] ou Youtube analytics[78]. D'autres le monétisent, tel que Kissmetric[79] ou Chartbeat[80], qui mesure l'attention des utilisateurs pour plus de 50 000 sites de média sur Internet. L'avantage de ces solutions est leur automatisation totale en temps réel sur un nombre d'éléments quasi illimité. La grande limite de ces nouvelles métriques est leur manque de validation académique, qui est très difficile à établir pour une personne n'ayant pas accès aux données des plates-formes associées.

La Triangulation : l'utilisation conjointe de différentes mesures

La mise au point de mesures ayant pour objet des entités de plus en plus abstraites et complexes demande une vigilance méthodologique accrue et un déploiement de moyens plus importants. Nombreuses de ces entités, telles que les émotions, l'état d'immersion ou l'expérience utilisateur ont pris une place centrale dans les domaines des IHMs (Hassenzahl & Tractinsky, 2006; Jennett et al., 2008; Norman, 2004a), des sciences de l'information (Fu, Su, & Yu, 2009; P. Zhang, Li, & Sun, 2006) et même du marketing (Erevelles, 1998; Fornerino et al., 2006) car elles constituent de bons indicateurs de la qualité d'un produit. Cependant, nous avons vu précédemment que ces construits sont plus enclin aux erreurs de mesures que les construits plus concrets. La combinaison d'approches différentes −mais complémentaires− est une piste prometteuse pour permettre l'évaluation de construit multidimensionnel et de réduire l'erreur de mesure associée (Churchill & Bardzell, 2007; Lai-Chong Law, 2011). Cette nouvelle

76 http://www.kaushik.net/avinash/best-social-media-metrics-conversation-amplification-applause-economic-value/
77 http://www.google.com/analytics/
78 https://www.youtube.com/analytics?o=U
79 https://www.kissmetrics.com/
80 https://chartbeat.com/

manière d'appréhender ce domaine est permise par les avancées de ces dernières décennies, tout aussi bien sur les plans techniques, informatiques et statistiques, que sur les plans méthodologiques, théoriques et épistémologiques (Michael Eid & Diener, 2006; Hetherington, 2000).

Avec plus de 4,200 citations, l'article de Campbell et Fiske « *Convergent and Discriminant Validation by the Multitrait Multi-method Matrix* » (1959) est l'un des premiers papiers permettant à la communauté de prendre conscience de l'intérêt d'utiliser plus d'une méthode pour s'assurer de la validité d'une mesure. Depuis lors, le champ de recherche s'est étendu, puis diversifié. Ainsi, il se retrouve de nos jours sous le nom de la triangulation (Webb, Campbell, Schwartz, & Sechrest, 1966), de l'approche multi-méthodes (Brannen, 1992), Multi-stratégies (Bryman, 2004), ou encore de méthode/méthodologie mixte (Creswell, 2003; Tashakkori & Teddlie, 1998). Ces diverses dénominations, bien que se fondant sur une pratique commune, c'est-à-dire l'utilisation de méthodes multiples pour étudier un phénomène commun (Denzin, 1978), répondent chacune à des intérêts particuliers et s'organisent selon différents niveaux.

Les objectifs des protocoles multi-méthodes

Greene (2005) donne cinq raisons qui justifient l'utilisation conjointe de différentes méthodes. Premièrement, l'obtention de résultats convergents, par la triangulation de plusieurs méthodes, augmentent leurs validités et leurs crédibilités respectives. Deuxièmement, l'utilisation des résultats d'une méthode peut permettre d'améliorer la qualité d'une autre. Troisièmement, l'utilisation de méthode complémentaire permet d'étendre les résultats obtenus par une méthode donnée et ainsi d'améliorer sa compréhension. Quatrièmement, cela permet de générer de nouvelles pistes de recherche, notamment en relevant des divergences entre les méthodes, qui demanderont une réconciliation via une analyse plus approfondie. Cinquièmement, cette approche permet de faire dialoguer des méthodes différentes, se fondant sur des valeurs et épistémologies particulières, multipliant ainsi nos manières d'appréhender le monde. Il n'est pas rare dans la pratique que plusieurs de ces buts s'entremêlent. Dans une approche plus extensive, Bryman (2008) référence seize justifications différentes, issues d'une revue de 232 cas d'étude utilisant des protocoles multi méthodes, dont les plus importantes sont l'amélioration de la validité, la complétude, l'échantillonnage, la compréhension et la diversité des vues. Enfin, de manière plus générale, en nous appuyant encore sur le dilemme de fidélité-bande passante de Cronbach (1960), ces objectifs peuvent être vus sous la forme d'un continuum, allant de la recherche de l'exhaustivité à la recherche de la validité. En effet, il est toujours possible, à partir d'un jeu de données similaires, de traiter chaque pièce individuellement (améliorer la complétude), ou de les agréger afin d'obtenir un indicateur plus fiable (améliorer la validité).

Amélioration de la complétude

La recherche de la complétude implique l'investigation d'un phénomène en utilisant une multitude d'éléments ou de méthodes, avec l'intention de se compléter et non de se répliquer. Elle peut prendre un certain nombre de formes, comme la génération d'hypothèses de recherche avant une étude ou l'explication des résultats par une autre. De plus, les résultats issus de plusieurs méthodes peuvent être juxtaposés pour élargir les insights et la compréhension d'un

phénomène (Olsen, 2004), chacune d'entre elles ne capturant qu'une partie limitée de la situation globale.

Ce genre d'approche est souvent utilisé dans les domaines peu ou pas explorés. Elle permet de générer une masse importante de données, qui peut aider les acteurs suivants à développer des approches plus rigoureuses et spécialisées. Cette approche est d'autant plus primordiale quand les d'entités traitées possèdent de multiples facettes, telles que les émotions, qui englobent des composants d'ordre physiologique, affectif, comportementale et cognitif (Brave & Nass, 2009). De même, nous ne pouvons pas saisir le fonctionnement de la communication humaine si nous ne considérons pas tous ses canaux d'informations, véhiculant des expressions faciales, des gestes, le ton de la voix et des mots (Paulmann, Jessen, & Kotz, 2009). Enfin, l'analyse des différentes composantes d'un phénomène permet de détecter les règles générales et individuelles des comportements à son égard. Par exemple, une personne qui grandit dans une culture où la fierté est une émotion très appréciée va l'exprimer explicitement (par exemple, aux États-Unis), tandis que d'autres élevés dans une culture où la fierté est importune la cacheront (par exemple, en Asie de l'Est) ; (Eid & Dienier, 2001). Ainsi, « *la combinaison des approches pour analyser les différences individuelles à partir de la covariation des différents éléments d'un phénomène multi-composant peut nous aider à comprendre les processus de régulation individuels et sociaux* » (Eid & Diener, 2006, p. 4)

Cependant, comme nous l'avons exposé précédemment, ouvrir le champ d'investigation lors d'une recherche a pour conséquence de diminuer la fiabilité des données récoltées. En effet, même dans le cadre d'une recherche expérimentale, où des garde-fous ont pourtant été élaborés afin de se prémunir des erreurs, l'augmentation du nombre de cas étudiés individuellement aura pour conséquence de multiplier les chances d'arriver à des conclusions incorrectes. Le chercheur peut être exposé ainsi à deux types d'erreurs statistiques: trouver un lien entre plusieurs variables alors qu'il n'y en a pas (l'erreur de type I, ou faux positif) et ne pas trouver de lien alors qu'il en existe un (l'erreur de type II, ou faux négatif). En effet, si nous choisissions une marge d'erreur acceptable de 5% pour un test, alors la chance de ne pas souffrir d'une erreur de type I est de 95%. Mais si nous voulons réaliser trois tests successifs de corrélation, la probabilité de ne réaliser aucune erreur de type I est de $1-(0.95)^3 = 1-0.857 = 14{,}3\%$[81]. La même démonstration s'applique aux erreurs de type II. Ainsi, si assez de tests de relations sont réalisés, il est potentiellement assuré qu'un certain nombre de ces tests apparaîtront statistiquement comme significatifs, car tout ensemble de données contenant un certain degré de perturbation aléatoire possédera en son sein des corrélations fortuites. Ce type d'erreur se rencontre également dans les études du type qualitatives, étant donné le nombre important de relations observées simultanément et la capacité inhérente de l'homme à voir de l'ordre dans le plus profond des chaos.

Amélioration de la précision

La recherche de la précision implique l'investigation d'un phénomène en utilisant une multitude d'éléments ou de méthodes dans le but d'en améliorer la crédibilité, la précision et la validité

[81] Erreur de type I $= 1 - (0.95)^n$, où *n* représente le nombre de test dans la même étude.

(Denzin, 1978; Jick, 1979; Webb et al., 1966). Cette approche fait souvent référence à la métaphore de la triangulation, qui vient de la navigation et de la stratégie militaire, et utilisent de multiples points de référence pour localiser la position exacte d'un objet. À ce sujet, McGraph et al. (1981) déclarent que « *c'est seulement quand nous avons des informations convergentes sur le même problème, acquises à partir de méthodes différentes, que nous pouvons parler d'accroissement de la connaissance* ». Denzin (1978) identifie quatre types de triangulation : l'utilisation de différentes personnes à différents endroits et à différents moments (triangulation de données) ; l'utilisation de plusieurs évaluateurs (triangulation des investigateurs), l'utilisation de plus d'un schéma théorique pour interpréter le phénomène (triangulation théorique) ; et l'utilisation de plus d'une méthode pour collecter les données (triangulation méthodologique). De plus, Denzin distingue la triangulation « intra-méthodes » et « inter-méthodes ». La triangulation « intra-méthodes » se réfère à l'utilisation de techniques multiples au sein d'une même méthode pour collecter et interpréter les données, en utilisant des items focalisés sur un même construit dans une enquête. La triangulation « inter-méthodes » se réfère à l'utilisation de plusieurs méthodes distinctes pour effectuer une validation croisée. Jick (1979) ajoute que la triangulation « intra-méthodes » implique essentiellement un contrôle de la consistance ou la fiabilité de la mesure, alors que la triangulation « inter-méthodes » teste son degré de validité externe.

Cependant, comme exposé précédemment, concentrer les efforts d'une recherche sur la création de quelques informations de qualité a l'inconvénient d'être coûteux en temps et en moyens déployés. De plus, le choix des données servant à la triangulation ne se fait pas sans de bonnes présomptions théoriques, car une agrégation inappropriée peut non seulement endommager la validité d'une mesure mais également dissimuler des patterns de données importants et conduire à des conclusions erronées (voir Schmitt, 2006).

Les niveaux de la triangulation

Les stratégies multi-méthodes peuvent s'organiser selon différents niveaux. Jick (1979) voit la triangulation comme un continuum allant d'une conception basique à une conception complexe. Nous présenterons ici trois niveaux : les multi-items, les multi-facettes et les multi-mesures.

Niveau multi-Items

Le niveau le plus bas est représenté par les multi-items ou le « multipoints ». Ce niveau représente la triangulation « intra-méthode » de Denzin (1978) et l'étage ordinaire de la théorie de test classique (Novick, 1966), centrale dans le domaine de la psychométrie. Ce niveau introduit les notions d'échantillonnage et de fiabilité. Dans les études quantitatives, l'échantillonnage consiste à générer un ensemble d'éléments ou d'indicateurs liés à un construit d'intérêt. Dans les études qualitatives, cela consiste à former une liste de thèmes liés à une matière spécifique. Un bon échantillonnage, conduit à une bonne validité de contenu. Cela souligne le fait que l'échantillon d'items rassemblé capte une bonne partie du spectre théorique associé à l'objet d'étude (Nunnally & Bernstein, 1994). Ainsi, deux types d'erreurs peuvent altérer la validité de contenu : l'exclusion d'indicateurs pertinents et l'inclusion d'indicateurs non pertinents. La meilleure façon d'éviter ces erreurs est de générer les items sur la base d'une revue exhaustive de la littérature et de juger sa représentativité par un comité d'experts

(Netemeyer, Bearden, & Sharma, 2003). Il est important de préciser qu'à ce niveau le set d'items généré reste dans le périmètre d'une méthode seulement.

L'échantillonnage est tout aussi important dans le cadre de la recherche de la complétude que de la précision. En effet, pour améliorer la complétude, la constitution d'un échantillon d'indicateurs variés permet d'étudier un phénomène sous de nombreux angles en même temps. De l'autre côté, la précision de la mesure augmente si elle s'assoit sur un grand échantillon d'observation. Cette approche, qui est discutée en détail par Cronbach et al. (1972), est utilisée largement dans les sciences naturelles et sociales. En effet, la fiabilité d'une mesure peut être augmentée par agrégation (Epstein, 1986; Steyer & Schmitt, 1990), qui suit directement la logique de la multi-détermination (Schmitt, 2006) : si la valeur de différents indicateurs est causée partiellement par un facteur commun et partiellement par des facteurs uniques, chacun d'entre eux est une mesure pauvre du facteur commun. Par contre, la moyenne des indicateurs réduit l'impact des facteurs uniques, alors que celui du facteur commun reste le même. De plus, plus le nombre de mesures est important, plus l'indicateur obtenu est stable, car l'erreur aléatoire de chaque mesure s'annule entre elles. Ce principe fait partie intégrante de la théorie du test classique et est la raison pour laquelle elle affirme que la fiabilité d'un test dépend de sa longueur (Lord & Novick, 1968). Avoir plus d'une mesure permet également le contrôle de sa qualité. Un vieil adage dit qu'un homme avec une montre sait l'heure qu'il est, un homme avec deux montres n'est jamais sûr. En d'autres mots, nous ne pouvons contrôler la fiabilité d'une mesure que si nous pouvons la confronter à une autre et constater si le résultat est (approximativement) similaire. Différents tests de fiabilité existent se basant sur ce principe : le test-retest, qui estime la stabilité des réponses aux items au cours d'une période de temps ; la consistance interne, qui estime le degré d'interrelation entre un ensemble d'item conçu pour mesurer un construit commun (souvent mesuré par le coefficient Alpha ; Cronbach, 1951) ; et l'accord inter-juges, qui correspond au degré d'homogénéité d'une évaluation commune à plusieurs juges, souvent mesuré par le coefficient Kappa (Cohen, 1960) ou par le coefficient de corrélation intraclasse (ICC ; Shrout & Fleiss, 1979). Pour augmenter la fiabilité de la mesure, on peut choisir d'écarter les items les moins fiables ou de les pondérer à partir de coefficients obtenus par une analyse factorielle (Pearson, 1901).

Toutefois une trop grande fiabilité, surtout pour des construits complexes, peut nuire à la validité de la mesure, car écarter tous les items un peu différents des autres conduit à une mesure partielle, qui n'évalue plus le construit dans son ensemble. Une autre limite de cette approche est l'utilisation d'une seule méthode, qui génère des erreurs systématiques cachées. En effet, «

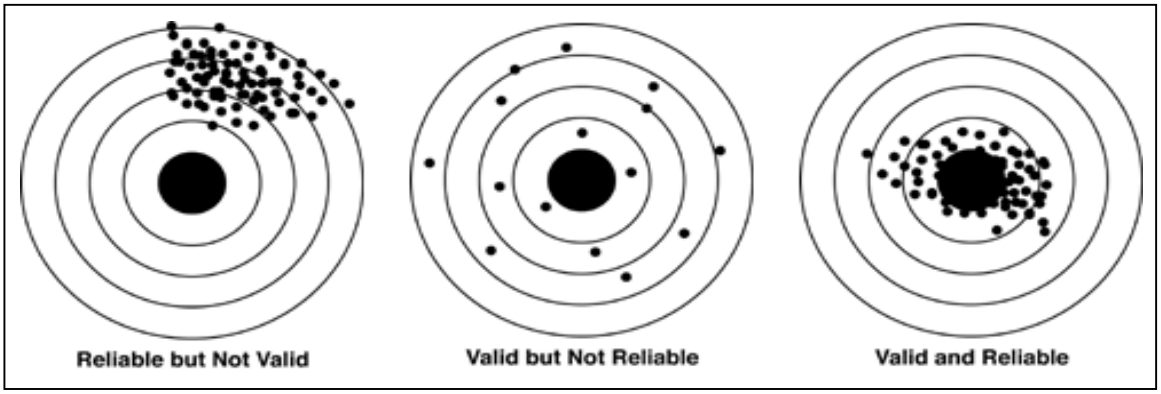


Figure 77 – Métaphore pour les erreurs de mesures affectant la fiabilité et la validité d'une évaluation

les observateurs se trompent en croyant que cinq variantes d'une même méthode génèrent cinq variétés distinctes de données triangulées, car les défauts qui surviennent en utilisant une même méthode restent ... » (Denzin, 1978). Une mesure peut ainsi être fiable, tout en n'étant point valide (ex : la cible 3 de la figure 77). Enfin, la théorie classique de la mesure, si elle peut estimer différents types d'erreurs, tels que l'erreur inter-juge ou instrumentale, ne peut le faire qu'une à la fois, entraînant une estimation incorrecte de la taille et de la source des erreurs.

Niveau multi-Facettes

Le niveau multi-facettes de la triangulation à pour philosophie de percevoir la mesure comme reposant sur une ensemble de facettes, dont chacune d'entre elles a un impact plus ou moins important sur la mesure en elle-même. L'évaluation multi-facettes étend ainsi le modèle psychométrique de base pour y inclure plus de variables (appelées facette), que les deux typiquement incluses (les personnes et les items), pour améliorer la complétude et la validité de la mesure. Elle est portée, entre autres, par la théorie de la généralisabilité (Cronbach et al., 1972; Cronbach et al., 1963) et les modèles de Rasch multi-facettes (Linacre, 1989).

Chaque facette est décrite par l'ensemble des conditions la constituant : items, tâches, occasion, évaluateur, etc. Dans la théorie de la généralisabilité, deux types de facettes peuvent être définis : les facettes de différenciation et les facettes d'instrumentation (ou de mesure). Les facettes de différenciation sont celles que l'on choisit d'étudier explicitement. Par exemple, un sociologue qui étudie les coutumes du mariage au cours du temps pourra prendre en compte lors de son analyse la facette de différenciation « époque ». Les facettes d'instrumentation sont celles qui serviront à contrôler la qualité de la recherche. Par exemple, une facette « couple mariée », composé d'un ensemble représentatif de couples permettra de bonne inférence sur la population étudiée. De même, une facette « investigateur », composée par de nombreux chercheurs indépendants, permettra de neutraliser les biais provenant des jugements idiosyncratiques. Enfin, la facette « méthode » composée d'un ensemble d'outils d'investigation, tels que les observations, les questionnaires ou encore les entretiens, permettront de contrebalancer les biais inhérents à chacune d'entre elles.

Par rapport à la théorie classique de la mesure, la théorie de la généralisabilité permet de s'interroger sur des entités autres que des individus. En effet, ce modèle permet d'étudier les caractéristiques de dispositifs conçus pour "mesurer" (= différencier) : items, objets, méthodes, etc.. Cette propriété est particulièrement adaptée au domaine des IHM, car il s'agit d'évaluer les systèmes numériques et non les individus. De plus, avec l'avènement de l'UX, nous pouvons étudier précisément la facette utilisateur (cachée dans les protocoles psychométriques classiques), qui, à cause de la subjectivité lors d'une interaction, pourrait contribuer grandement à la variabilité de l'évaluation. Par exemple, nous pouvons imaginer un protocole de mesures avec une facette de différenciation « interface » et trois facettes de mesure « utilisateur », « item » et « méthode ». Cela nous permettrait d'estimer l'impact de chaque facette et leurs interactions sur la fiabilité de la mesure, et ainsi nous permettre de nous concentrer sur celle qui mérite le plus de travail.

Niveau multi-mesure

Le niveau multi-mesure de la triangulation consiste à agréger des objets obtenus à partir de méthode différentes pour contrebalancer les biais systématiques inhérents à chacune d'entre elles. Cela permet une augmentation de la validité de la mesure. Par exemple, il peut s'agir de combiner des données issues d'un questionnaire, de mesures physiologiques et de données faciales pour estimer l'état émotionnel d'un individu. C'est un sujet important dans le domaine de l'UX, dont la problématique principale est de savoir comment agréger des données de nature si disparates (Churchill & Bardzell, 2007; Law, 2011). En effet, la majorité des études actuelles recueillent plusieurs types de mesures le font pour les confronter et non pour les combiner, car ils estiment ne pas disposer d'assez de données sur leurs qualités respectives pour le faire. Par exemple, Ortiz De Guinea et al. (2013), en utilisant la Matrice Multi-Trait Multi-Methodes (MTMM) de Campbell et Fiske (1959), étudient les liens corrélationnels entre les méthodes subjectives et neurologiques sur la mesure de l'engagement, de l'excitation émotionnel et de la charge cognitive. Ces protocoles multi-mesures ont été mis en place pour étudier divers construits tels que l'engagement émotionnel (Jenkins, Brown, & Rutterford, 2009; Kukolja, Popović, Horvat, Kovač, & Ćosić, 2014; Tep, Dufresne, Saulnier, & Archambault, 2008; Siefert et al., 2009), ludique (Bianchi-berthouze, 2013; Cui & Rau, 2012; C Jennett et al., 2008; Matias Kivikangas, Nacke, & Ravaja, 2011; Lennart E. Nacke et al., 2010; Peter & Herbon, 2006), attentionel (Sundar et al., 2010) ou le flow (Ivory & Magee, 2009). Des méthodes de combinaison de mesures ont été mises au point dans le paradigme de l'utilisabilité (Sauro & Kindlund, 2005; Tullis & Albert, 2008), mais ne sont encore que peu exploitées dans le domaine de l'UX. Par contre, la fusion d'indicateurs, par des modèles puissants issues du Machine Learning, a été largement exploitée ces dernières années (Asteriadis et al., 2009; Atrey, Hossain, El Saddik, & Kankanhalli, 2010; Bailenson et al., 2008; D'Mello & Graesser, 2010; Gilroy, Cavazza, & Vervondel, 2011; Thiran et al., 2010). On notera toutefois que les études réalisées avec ces dispositifs mathématiques complexes déploient des efforts disproportionnés pour la mise au point de leur modèle en comparaison à ceux qui sont déployés pour conceptualiser les construits mesurés. Ce qui conduit à la mise au point d'un outillage extrêmement compliqué au service de la mesure d'un concept flou. Un compromis théorique et méthodologique serait donc à mettre en place.

L'approche multimodale est particulièrement intéressante dans le cadre de la mesure de l'UX. Le concept de modalité fait référence aux différents modes d'accès dont nous disposons pour accéder à l'information. Comme nos sens, les modalités sont souvent complémentaires, et, utilisées ensemble, elles nous permettent de capter convenablement un message, même en présence d'éléments perturbateurs ou de signaux en partie contradictoires. Par exemple, la reconnaissance des émotions chez les humains est instinctivement multimodale. En nous basant sur la totalité des signaux émotionnels que nous percevons chez un individu, tels que ses expressions faciales, sa voix, son langage corporel, et même certaines de ses réactions physiologiques (comme le rougissement), nous sommes capables de déceler son état affectif en un instant, même ceux réprimés partiellement (Clay, 2009). Ainsi, une approche multimodale de l'évaluation de l'expérience utilisateur, en se basant sur des modes d'accès à l'information complémentaires, est essentiel pour apprécier totalement des construits aussi complexes.

La combinaison de données de nature très différentes est une opération délicate, qui consiste parfois à se demander comment mélanger des pommes et des oranges. Elle peut s'appuyer sur des modèles mathématiques très puissants, tels que la fusion multimodale, qui peut combiner en temps réel des données d'ordre physiologique (Gilroy et al., 2011; Karpouzis, 2011) ou multimédia (Atrey et al., 2010). Des techniques plus simples sont également utilisées pour combiner des indicateurs discrets (non continus) en indicateurs composites. Par exemple, dans le domaine de la géopolitique, les indices sont créés sur ce principe. L'indice de développement technologique combine ainsi huit facteurs, qui sont (i) le nombre de brevets, (ii) la réception de droits d'auteur, (iii) le nombre d'accès à Internet, (iv) les exportations de produits de haute et moyenne technologie, (v) le nombre de lignes téléphoniques, (vi) la consommation d'électricité, (vii) le nombre moyen d'années de scolarité, et (viii) le taux de scolarisation des étudiants du supérieur en sciences, mathématiques et en ingénierie (Desai, Fukuda-Parr, Johansson, & Sagasti, 2002). Autre exemple, l'indice de développement humain (IDH) est un indice statistique composite créé par les Nations Unies, en 1990, et dont le calcul a été modifié en 2011[82]. Il est calculé par la moyenne géométrique de trois indices quantifiant respectivement la longévité, le niveau d'éducation et le niveau de revenu. Chacun de ces trois indices suivent également une règle de calcul bien précise (par exemple, le niveau de revenu est mesuré à partir du logarithme décimal du PIB par habitant en parité de pouvoir d'achat). Nous voyons donc que ces indicateurs composites se construisent à partir d'une connaissance approfondie du domaine d'étude, et, à chaque étape, des choix subjectifs doivent être réalisés. Un guide a ainsi été mis au point sur la commande de l'OCDE, et nous renseigne, à chacun des stades de la construction de l'indicateur composite, des précautions à prendre et les contrôles à effectuer (Nardo et al., 2005). Il s'articule en dix étapes, dont les six plus importantes sont :

1. **LE CADRAGE THEORIQUE :** dont l'objectif est une compréhension claire du phénomène multidimensionnel à mesurer. Il constitue la base pour la sélection et la combinaison de variables pertinentes à la construction de l'indicateur composite
2. **LA SELECTION DES INDICATEURS :** Basée sur la mesurabilité et la pertinence des indicateurs par rapport au phénomène mesuré, en estimant la force et la faiblesse de chacun
3. **L'IMPUTATION DES DONNEES MANQUANTES :** Indispensable pour travailler sur un ensemble complet de données. Diverses méthodes existent, telles que l'imputation simple, l'imputation multiple, l'imputation par moyenne inconditionnelle, l'imputation par régression, ou l'imputation par « *Espérance-maximisation* » (EM).
4. **L'ANALYSE MULTIVARIEE :** Pour étudier la structure globale des données, évaluer sa pertinence, et orienter les choix méthodologiques ultérieurs
5. **LA NORMALISATION :** Procédure utilisée pour rendre les variables comparables, en choisissant des techniques respectant à la fois le cadre théorique et la nature des données. Diverses méthodes existent, tels que l'ordonnancement (t), la standardisation (ou z-scores), la distance à une référence, le Min-Max, etc.

[82] Tel qu'indiqué dans les notes techniques du rapport sur le développement humain de 2011, disponible ici : http://hdr.undp.org/en/media/HDR_2011_FR_TechNotes.pdf

6. **LA PONDERATION ET L'AGREGATION** : Basées sur des procédures respectant la structure du modèle théorique et optimisant sa validité. La pondération peut se baser sur la méthode des poids égaux (EW / Equal Weighting), l'analyse factorielle (FA) / analyse en composantes principales (PCA), ou encore sur des procédures d'optimisation de la fiabilité. L'agrégation peut être linéaire ou géométrique

Ce processus peut être complété par des analyses d'incertitude et de sensibilité pour tester la robustesse de l'indicateur composite (Cherchye et al., 2008). L'ensemble de cette procédure peut être adaptée à la création de mesures composites de l'expérience utilisateur selon une approche multimodale, c'est-à-dire en combinant des indicateurs d'ordre subjectif, comportementale et physiologique.

Une nouvelle épistémologie : le « critical multiplism »

En conclusion de cette partie, nous voyons que la triangulation peut être réalisée à tous les niveaux, pour augmenter la précision, la diversité, la complétude ou encore la validité de l'évaluation. Elle peut se faire au niveau de l'élément de base, l'item, se poursuivre par la multiplication des facettes qui composent un protocole de mesure, ou encore par l'utilisation de plusieurs méthodes conjointes ne partageant pas les mêmes biais. À un niveau encore plus élevé, la multiplication des études sur un même sujet, synthétisé dans un état de l'art ou dans une méta-étude, relèvera de la multi-stratégie (Carmines & Woods, 2011).

Toute approche seule, qu'elle repose sur un outil de mesure, une méthode, une théorie, possède ses biais et ses points d'ombres. Utiliser une approche seule occultera ses biais de sorte qu'ils ne puissent pas être reconnus. C'est là où l'épistémologie du « *critical multiplism* » peut apporter un remède (Patry, 2006). Ce mouvement milite pour l'utilisation de plusieurs approches pour répondre à une question particulière, mais pas dans le sens d'un activisme aveugle du type « *tout convient* », mais au contraire dans un sens critique et systématique. En d'autres termes,, que ce n'est pas le nombre d'options mis en œuvre qui importe, mais plutôt « *que les différentes options sélectionnées possèdent des biais qui opèrent dans des directions différentes* » (Shadish, 1986). Ainsi, cette approche atteint son but quand la diversité injectée à la résolution du problème particulier souligne sa complexité et révèle les biais de chacune des techniques isolés.

PROBLEMATIQUE

Nous avons vu que le domaine des applications de communications a connu dernièrement une grande effervescence. Parmi ces évolutions, les plus remarquables se caractérisent par la convergence des types de communications et de contenus ; un mode de gestion des applications plus communautaire et participatif et à la définition de produits, non plus seulement utiles et utilisables, mais également satisfaisant une expérience utilisateur holistique. La diversité des médiums émergeants, brisant le cloisonnement taxonomique précédemment établi, entrave l'application des anciennes méthodes d'évaluation basées sur les connaissances expertes et des métriques spécialisées. Ce bouleversement a conduit les chercheurs à suivre un nouveau paradigme d'évaluation, basé sur l'expérience utilisateur. Ce courant se différencie fondamentalement de ses prédécesseurs par son caractère holistique, subjectif et positif. Un grand nombre de méthodes de mesures ont ainsi été mises au point afin de mesurer cette expérience utilisateur, mais demeurent immatures, voire contradictoires entre elles.

De plus, nous avons vu que deux autres grandes transformations du domaine ont modifié les méthodologies privilégiées pour la conception et l'évaluation des systèmes numériques. D'une part, l'accélération du renouvellement de l'offre, par la mise en place d'une économie de l'innovation, a concentré les efforts en amont de la conception, en s'appuyant sur des méthodes qualitatives, alliant créativité et anticipation des besoins. D'autre part, la mise en ligne précoce des produits, couplée à des mécanismes internes de recueil du feedback utilisateurs (majoritairement quantitatifs), a permis de poursuivre à faible coût l'adaptation des services en fonction des exigences des utilisateurs ciblés.

Néanmoins, dans le domaine de l'évaluation, aucune des approches actuelles n'a pu maîtriser toutes les étapes nécessaires à la création d'une mesure robuste de l'expérience utilisateur. Nous avons vu que l'ergonomie a les moyens d'y parvenir, en s'appuyant sur les outils actuels (statistiques, techniques et informatiques) et le sérieux de ses méthodologies de modélisation des construits complexes. Ce processus s'appuie sur une phase de conceptualisation solide, une opérationnalisation à travers différentes mesures, puis par une validation. De plus, dans le contexte de la mesure d'un construit abstrait comme l'UX, nous pensons qu'il convient d'ajouter une dernière étape, celle de la triangulation, en s'appuyant sur un large spectre d'indicateurs, d'ordre physiologiques, comportementaux et auto-rapportés. Ces mesures composites seront testées dans des cas d'applications réelles et selon une complexification croissante des procédures et traitements statistiques associés.

L'accent sera particulièrement mis sur le cas d'étude de la mesure de l'immersion dans une situation de vidéo-conférence, qui correspond à un axe stratégique de recherche propre à Alcatel-Lucent. Cette thèse se situant dans un cadre industriel, cette étude conduira à l'établissement de recommandations pour mettre au point des protocoles d'évaluation multimodaux visant un rapport optimal entre la qualité et l'effort de mesure.

ÉTUDES

PRESENTATION DES ETUDES

Cette thèse se compose de trois études, dont la complexité théorique, méthodique et statistique croît graduellement. De plus, les leçons tirées de chaque étude sont injectées dans la suivante.

Etude 1 – Mesure multifacette de la pertinence des recommandations de films

La première étude a été réalisée en support d'un autre chercheur des Bells Labs. Elle a consisté à évaluer la pertinence d'un algorithme de recommandation de films face à son concurrent. L'étude utilise une stratégie d'évaluation multifacette, basée sur la théorie de la généralisabilité de Cronbach. L'objectif de cette étude a été de se familiariser avec ce type de protocole de validation de mesure pour être utilisée dans les suivantes. Le protocole de mesure a pris en compte une facette de différenciation (« *Algorithme* », à deux modalités) et quatre facettes d'erreur (« *Juge* », « *Recommandation* », « *Ordre* » et « *Film* »). L'utilisation d'un modèle multifacette nous a permis de mettre à jour des enseignements transférables dans le domaine de l'utilisabilité et de l'UX.

Etude 2 – Mesure multimodale et à distance de l'utilisabilité appliquée aux sites universitaires

La deuxième étude a été élaborée afin de tester la pertinence des modèles d'évaluation multimodale. Elle a consisté à évaluer l'utilisabilité d'un certain nombre de sites universitaire grâce à un logiciel de test utilisateur à distance (Evalyzer). Pour ce faire, nous avons demandé à une trentaine d'étudiants de Metz de réaliser trois tâches de recherche d'informations et de remplir un questionnaire SUS (System Usability Scale) pour chacun des 5 sites universitaires étudiés. De plus, de nombreuses données comportementales ont été capturées, telles que le temps de réalisation des tâches, la distance à la souris, le nombre de clics, le nombre d'URL visités, le taux de revisite et la réussite des tâches. Ces données ont été ensuite standardisées et combinées pour mettre au point différents indicateurs composites de l'utilisabilité. Le protocole de validation multifacette comprend une facette de différenciation (« *Site universitaire* », à cinq modalités) et trois facettes d'erreurs (« *Etudiant* », « *Type de mesures* » et « Items »). Cette étude a permis de tester la pertinence des protocoles multifacettes dans le domaine particulier des interactions homme-machine et dans le contexte de l'évaluation à distance. De plus, elle a permis de démontrer l'intérêt de mettre au point des indicateurs composites basés sur des données récupérées automatiquement par le système.

Etude 3 – Mesure de l'immersion multimodale dans le cadre d'application de vidéoconférence

La troisième étude a eu pour objectif de valider un protocole complet de mesures multimodales de l'immersion dans le contexte spécifique d'une activité de vidéoconférence. La première expérience a été réalisée sur 60 utilisateurs et a permis de recueillir des données de type physiologiques (conductances de la peau, expressions faciales), comportementales (verbalisations écrites) et auto-rapportées (questionnaires). Elle a été complétée par une autre expérience sur 103 utilisateurs, ce qui a permis de récupérer d'autres types d'indicateurs (questionnaire, expression faciale, conductance de la peau, rythme cardiaque, comportement oculaire) et de fournir des données suffisantes pour tester la pertinence des protocoles de

mesures multimodales dans le domaine particulier de l'immersion. Le protocole prend en compte une facette de différenciation (« *Interface* » à trois modalités) et quatre facettes d'erreurs (« *Utilisateur* », « *Présentation* », « *Mesure* » et « *Items* »). Dans un premier temps, cette analyse a permis d'identifier les indicateurs les plus pertinents pour mesurer l'immersion, grâce à l'étude de leurs qualités psychométriques respectives. L'analyse des résultats s'est ensuite poursuivie par la triangulation multimodale des indicateurs d'immersion jugés pertinents et leur validation.

ETUDE 1 – MESURE MULTIFACETTE DE LA PERTINENCE DES RECOMMANDATIONS DE FILMS DANS LE CADRE DE L'EVALUATION D'ALGORITHME DE PLATEFORMES DE RECHERCHE EXPLORATOIRE

Introduction

Parmi les sujets d'étude des Bells Labs existants, le web sémantique a été un axe de recherche relativement important. Cette approche vise à lier et à structurer l'information sur Internet pour accéder simplement à la connaissance qu'elle contient déjà. L'utilisation des liens entre ces données a été utilisée par des chercheurs au sein des Bell Labs pour améliorer les performances des plates-formes de recherche exploratoire actuelles (Marie, 2014). C'est dans ce cadre que j'ai été sollicité pour évaluer l'algorithme d'activation par propagation sémantique (basé sur DBpedia[83]). L'évaluation des systèmes de recherche exploratoire étant complexe (M. L. Wilson, Schraefel, & White, 2009), l'objectif de cette expérimentation fut de tester dans un premier temps la qualité des recommandations engendrées par l'algorithme, face à celles de la concurrence, sans se préoccuper de l'interface (Figure 78).

Figure 78 – Page de résultat de l'interface Discovery Hub (tirée de Marie, Gandon, Ribière, & Rodio, 2013)

Les résultats ont été publiés par ailleurs (Marie, Gandon, Ribière, & Rodio, 2013) et ne constituent pas le socle de cette étude. Il s'agit plutôt ici de présenter une réflexion critique de la méthodologie de validation utilisée et de proposer des pistes pour sa réutilisation dans le domaine de la mesure de l'UX.

Les méthodologies d'évaluation des plates-formes de recherche exploratoire

De nombreux chercheurs constatent que les évaluations dans le domaine de la recherche exploratoire souffrent d'un manque de standardisation (Leskovec & Faloutsos, 2006; Waitelonis & Sack, 2012). Plusieurs facteurs sont à l'origine de ce constat.

Premièrement, si l'on s'en tient au seul domaine des systèmes de recommandation, on voit naître un enrichissement des critères d'évaluation (Pu, Chen, & Hu, 2012). En effet, la recherche dans ce domaine se concentrait autrefois essentiellement sur la performance des

[83] DBpedia est une initiative de recherche publique visant à publier des données structurées extraites de Wikipédia et libre d'utilisation

algorithmes (Adomavicius & Tuzhilin, 2005; Herlocker, Konstan, Terveen, & Riedl, 2004), et plus spécifiquement encore sur leur précision de prédiction objective (Sarwar, Karypis, Konstan, & Riedl, 2001). Aujourd'hui, on reconnaît que la précision de la recommandation seule n'est plus suffisante pour satisfaire pleinement les utilisateurs, les encourager à réutiliser le service ou les persuader d'acheter (Herlocker, Konstan, & Riedl, 2000; Knijnenburg, Willemsen, Gantner, Soncu, & Newell, 2012; Konstan & Riedl, 2012; McNee, Riedl, & Konstan, 2006). Pu et al. (2012), propose de classer l'ensemble de ces indicateurs en quatre catégories : (i) la qualité générale des items proposés, telles que leur précision, nouveauté, attractivité ou diversité ; (ii) la facilité d'élicitation des préférences ou des processus de révision ; (iii) l'adéquation de la disposition et de la dénomination de la zone de recommandation ; (iv) l'habilité du système à assister et à améliorer la qualité des décisions utilisateurs ; et (v) la confiance qu'inspire les items recommandés aux utilisateurs.

Le deuxième facteur nuisant à la cohérence des évaluations est la méconnaissance de certains facteurs, qui, s'ils ne sont pas contrôlés, peuvent invalider notre capacité à comparer les résultats des études entre elles. Ces facteurs, souvent cachés, concernent les tâches-utilisateurs évaluées, le type d'analyse choisi, les échantillons de données testés, et les méthodes de mesure utilisées pour estimer la qualité de prédiction (Herlocker et al., 2004). Un algorithme pourra ainsi montrer des performances variables, à cause de la spécificité d'un domaine (ex : musiques, films, personnalités, …) ou des caractéristiques de l'échantillon de données testés, sans que ces spécificités ne soit contrôlées ou explicitées dans l'étude. Ainsi, les chercheurs qui veulent comparer quantitativement la précision de différents systèmes de recommandation doivent répondre à une série déterminante de questions, d'ordre psychométrique :

- Les résultats obtenus avec la métrique choisie peuvent-ils être comparés avec les autres études publiées dans le domaine ?
- Les hypothèses sur lesquelles se base l'utilisation de cette métrique sont-elles fondées ?
- Sera-t-elle assez sensible pour détecter des différences existant dans les données ?
- De quelle taille doit être une différence pour être détectée de manière significative par le protocole de test utilisé ?

Des réponses substantielles à ces questions n'ont pas encore été adressées dans la littérature publiée à ce jour dans le domaine. Elles sont directement liées à la variabilité des échantillons de données utilisées et aux positions idiosyncrasiques des utilisateurs participant aux études (Herlocker et al., 2004). Il est possible d'apporter des réponses élégantes à ces questions en utilisant la théorie de la généralisabilité (Cronbach et al., 1963).

La théorie de la généralisabilité

La théorie de la généralisabilité (G) est une théorie statistique utilisée pour évaluer la fiabilité d'un protocole de mesure. Elle permet de modéliser la précision d'une mesure en identifiant les sources potentielles de variabilité et en estimant leur magnitude. Alors que les théories classiques de la mesure (Novick, 1966) se basent sur des techniques statistiques d'ordre corrélationnel, la «G» théorie utilise l'analyse de la variance (ANOVA) pour partitionner la variance totale en un certain nombre de sources explicatives. Néanmoins, elle ne contredit pas l'approche classique, mais l'étend, par une unification des conceptions et des techniques

présentées classiquement de manière disparate (stabilité, équivalence, consistance interne, validité, consistance inter juges, …). Le fait que de nombreuses sources d'erreurs de mesure peuvent être incorporées simultanément dans un même modèle de mesure, tout en étant quantifiées séparément, permet l'exploration de divers plans alternatifs d'échantillonnage et d'en visualiser les effets. De ce fait, la théorie de la généralisabilité (G) est un outil puissant pour évaluer et concevoir des protocoles de mesure. C'est pourquoi elle est mise en avant dans les standards des test psychologiques et éducationnels, formalisée par l'APA (American Psychological Association), l'AERA (American Educational Research Association) et le NCME (National Council on Measurement in Education) ; (AERA, APA, NCME, 2002).

La théorie a subi deux étapes dans son évolution, ce qui l'a conduit à grandement diversifier ses applications. À l'origine, la théorie se situait implicitement à l'intérieur du cadre familier de la théorie du test classique, c'est-à-dire dans un cadre où l'individu (le patient, l'étudiant, etc.) était considéré comme l'objet de la mesure et dont l'objectif était de les différencier aussi précisément que possible. La principale exigence était donc de vérifier que l'instrument utilisé, le test ou le questionnaire, pouvait produire des mesures fiables de la position relative des individus sur l'échelle de mesure proposée, en dépit de l'influence perturbatrice des éléments de l'instrument de mesure elle-même (la part de variabilité aléatoire issue des tests ou des items d'un questionnaire). À partir des années 1970-1980, un élargissement de ses applications potentielles fut identifié, en se basant sur la symétrie inhérente au modèle ANOVA, qui n'était pas encore totalement exploité par la G théorie. En reconnaissant que n'importe quel facteur dans une conception factorielle pouvait devenir l'objet de la mesure, il était possible d'évaluer également des procédures de recherche ou des instruments de mesure à la place des individus (Cardinet, Tourneur, & Allal, 1976; Gillmore, Kane, & Naccarato, 1978). Même si, jusqu'à présent, cette propriété a surtout été exploitée dans le domaine de la psychologie ou de l'éducation, son application dans le domaine de l'évaluation des systèmes numériques est pertinente, dans le sens où la mise au point d'un protocole d'évaluation dans ce domaine vise à discriminer au mieux les applications entre elles, et non les individus qui les utilisent.

En résumé, l'objectif principal d'une étude de généralisabilité (ou étude G) est d'évaluer les caractéristiques d'un protocole de mesure donné et d'estimer sa précision. À cette fin, les différentes sources d'erreurs de mesures doivent être identifiées et quantifiées. Leurs importances relatives sont indiquées par la taille de leurs composants de variances associés, en utilisant une ANOVA classique. L'estimation de ces composants permet alors de calculer les erreurs de mesure et les coefficients G associés. L'étude G suit une série d'étapes qui conditionnent son protocole d'évaluation. Elles sont au nombre de cinq : (i) la mise en place d'un plan d'observation, (ii) l'échantillonnage des facettes, (iii) la sélection des objets de mesure, suivies de (iv) l'estimation des coefficients G et de l'erreur standard de mesure (étude G), et (v) la mise en place d'études d'optimisation (étude D). Le contenu de ces étapes est présenté brièvement dans les parties suivantes. La formalisation mathématique de la théorie est contenue à son minimum, mais assume toutefois que le lecteur possède quelques notions dans le domaine de l'analyse de la variance. Les fondations mathématiques de la méthode peuvent être consultées dans le texte fondateur de Cronbach et collaborateurs (Cronbach et al., 1972) et ses développements les plus récents dans l'ouvrage de Robert Brennan (2001).

Plan d'observation

Dans la théorie G, les sources de variations sont appelées « *facettes* » et sont similaires aux « *facteurs* » utilisés dans l'analyse de la variance. Elles peuvent se composer d'individus, de juges, d'items de questionnaire, d'instants d'évaluation, de configurations de test et bien d'autres possibilités encore. Quand une étude G est élaborée, les premières choses à identifier sont les facettes qui influent la précision de la mesure, ainsi que leurs relations. Par exemple, dans un test où l'on souhaite évaluer l'habilité cognitive de nombreuses personnes à partir d'un questionnaire, nous pouvons identifier la facette « *Personne* », qui contient toutes les personnes visées par le test, et la facette « *Item* » qui contient toutes les questions du questionnaire d'habilité cognitive. Contrairement à la théorie classique de la mesure, la G Théorie permet de tester en même temps la fiabilité d'un protocole possédant plus de deux facettes.

Il s'agit ensuite de spécifier les relations entre les facettes, qui peuvent être soit croisées, soit imbriquées. On dit que deux facettes sont croisées quand tous les niveaux d'une facette sont combinés avec tous les niveaux d'une autre. Par exemple, si toutes les personnes répondent à tous les items d'un questionnaire, alors les facettes « *Personne* » et « *Item* » sont croisées : cela est symbolisé par P x I, ou plus simplement par PI. Le concept de croisement peut s'étendre naturellement à des cas de plus de deux facettes. De même, on parle de facette imbriquée quand chaque niveau d'une facette est associé avec un et seulement un niveau de l'autre facette. Par exemple, si cinq juges évaluent la performance artistique d'un groupe d'athlètes et que cinq autres juges évaluent la performance d'un autre groupe, si aucun des juges n'évalue tous les groupes d'athlètes à la fois, alors on dit que la facette « *Juge* » est imbriquée dans la facette « Groupe » (que l'on note J(G) ou J : G). Le choix stratégique du croisement ou de l'imbrication des facettes repose sur divers facteurs. En effet, un plan entièrement croisé a l'avantage d'être plus robuste, car les biais systématiques de mesures de chaque facette, étant répercutés sur tous les éléments du protocole, s'annulent. Par exemple, si un juge très sévère (ou très clément) juge tous les groupes d'athlètes, alors les biais de la mesure s'annuleront ; au contraire, si ce même juge n'évalue qu'un seul groupe d'athlètes et qu'un autre juge plus clément en évalue un autre, alors un des groupes sera jugé plus défavorablement que les autres (biais de sévérité). De plus, les plans croisés ont pour avantage de fournir un plus grand nombre d'informations sur les sources d'erreurs issues des facettes et de leurs interactions. D'un autre côté, pour des raisons pratiques ou logistiques, il est parfois préférable d'imbriquer certaines facettes entre elles. Par exemple, si le nombre de conditions à tester devient trop important pour une même personne, ou si des biais d'apprentissage apparaissent, il vaut mieux séparer l'évaluation des conditions entre plusieurs groupes.

Chaque facette étant une source d'erreurs potentielle, l'étude G aura pour objectif de quantifier la charge d'erreur pour chacune d'entre elles, ainsi que leurs interactions. Dans le cas du test d'habilité P x I, la variance totale (σ^2_{T}) peut ainsi se décomposer en un certain nombre de composants : la variance des personnes (σ^2_P), la variance des items (σ^2_I) et la variance de l'interaction item-personne ($\sigma^2_{PI,e}$), ce dernier étant confondu avec la variance résiduelle :

$$\sigma^2_{\mathrm{T}} = \sigma^2_P + \sigma^2_I + \sigma^2_{PI,\mathrm{e}}$$

Une manière plus intuitive d'illustrer cette décomposition de la variance tout en visualisant le plan d'observation est de construire un diagramme de partition de variance (Cronbach et al., 1972, p. 37), qui s'inspire des diagrammes de Venn (Figure 79).

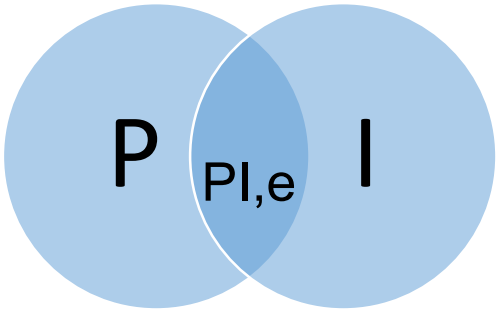


Figure 79 - Diagramme de partition de la variance pour le modèle P x I

Plan d'échantillonnage

Connaître la structure des données à partir d'un plan d'observation ne suffit pas pour estimer les composants de la variance et exploiter les bases théoriques de la généralisabilité. Il y a une autre caractéristique des données qu'il convient de définir : son plan d'échantillonnage.

Une facette peut être « *fixe* », « *aléatoire/infinie* » ou « *aléatoire finie* ». La taille de la population de la plupart des facettes étant généralement très importante, elle peut être considérée, d'un point de vue pragmatique, comme infinie. La population de la facette est alors appelée « *univers* ». Les utilisateurs et les items[84] sont des exemples d'entités que l'on ne peut qu'observer à partir d'un échantillon tiré d'une population bien plus grande et considérée comme infinie. Une facette est dite « *aléatoire* » quand les niveaux pris en considération (le nombre d'éléments de la facette utilisée dans l'étude) sont sélectionnés aléatoirement à partir de son univers d'appartenance. Si l'échantillon est tiré aléatoirement d'une population finie, on parle de facette « *aléatoire finie* ». Une facette est dite « *fixe* » quand tous ses niveaux sont représentés dans l'ensemble de données et qu'il n'y a pas eu, de fait, de procédure d'échantillonnage. Les facettes fixes « *naturelles* » sont rares. Il y a par exemple la facette « *sexe* » (avec deux niveaux : homme et femme), ou la facette « *région* ». La plupart des facettes fixes sont des facettes infinies qui ont été artificiellement fixées. En effet, l'avantage de fixer une facette est d'augmenter la fiabilité de la mesure, car cela élimine les fluctuations aléatoires dues à la procédure d'échantillonnage. Néanmoins, cela implique de réutiliser exactement les mêmes éléments de la facette d'un test à l'autre. Par exemple, une facette à 15 niveaux représentant les items d'un questionnaire peut être fixée, et exigera alors que tous les tests suivants utilisent le même questionnaire afin de pouvoir comparer les résultats des études avec la même fiabilité. Certaines facettes sont plus dures à fixer que d'autres. Par exemple, fixer une facette constituée de juges peut être un atout indéniable pour la fiabilité d'un test, mais nécessiterait alors d'utiliser toujours les mêmes juges, ce qui est très difficile en pratique. Une procédure de mesure peut contenir simultanément des facettes fixes et aléatoires : ce sont des modèles « *mixtes* ».

Le type d'échantillonnage d'une facette a des répercussions sur l'algorithme utilisé pour estimer les composants de variances et le coefficient G. C'est pourquoi cette étape doit être effectuée judicieusement.

Plan de mesure

L'identification des facettes, de la nature de leurs relations et de leurs types d'échantillonnage, permet l'estimation de leurs composants de variance respectifs à partir d'une ANOVA. Pour procéder à une

[84] Les items d'un test sont considérés comme des éléments tirés d'une facette infinie car ils sont issus d'une opérationnalisation particulière et contingente ayant pour objectif de mesurer un trait ou une habilité

analyse de la généralisabilité, il faudra également identifier « *l'objet d'étude* » et le type de mesure visé.

L'objet d'étude détermine quelle sera la facette de différenciation. Selon le principe de symétrie, toute facette peut être l'objet de l'étude. Dans notre exemple, cela pourrait être indifféremment les juges que les athlètes. Dans le premier cas, si on sélectionne comme objet d'étude les juges, on testera si notre modèle de test permet de discriminer avec précision leurs différents niveaux de sévérité ou de clémence. Si l'on choisit comme objet d'étude les athlètes, on testera la précision du modèle à discriminer leurs niveaux respectifs de performance. Après la sélection d'une facette de différenciation, les facettes restantes sont considérées comme des facettes d'instrumentation ou de mesure. Les facettes de différenciation génèrent de la « *variance de différenciation* », qui est similaire à la variance du « *score vrai* » dans la théorie de test classique. Elle correspond à la part de variance qui « *explique* » les différences effectives entre les entités mesurées. Toutes choses égales par ailleurs, plus la variance de différenciation est grande, plus la mesure sera précise. De l'autre côté, les facettes restantes produisent de la variance d'instrumentation, ou encore, selon les termes de la théorie classique de la mesure, de la variance « *d'erreur* ». En somme, de la variance qui produit du bruit se répercutant sur la précision de la mesure. Toutes choses égales par ailleurs, plus la variance de différenciation est petite, plus la mesure sera précise. On notera toutefois que les facettes fixes produisent moins de bruit du point de vue de la généralisabilité, car ils se passent d'échantillonnage. On peut constater ainsi que la qualité d'une mesure dépend de la taille de la variance des facettes de différenciation ET des facettes d'instrumentation. Il ne suffit pas que les facettes d'instrumentation génèrent peu de bruit pour que l'on puisse dire que la mesure est sûre. Le contraire est également vrai. Par exemple, un thermomètre ayant une précision de plus ou moins cent degrés ne peut pas être considéré comme un mauvais instrument de mesure en soi, car cette affirmation dépend des objets à mesurer. En effet, avec une facette de différenciation dont l'univers est composé de soleil (dont la température varie 3000 à 50000 degrés), une variation de plus ou moins cent degrés dans la mesure est négligeable ; Cela est, au contraire, rédhibitoire si ce même thermomètre est utilisé pour discriminer la température sur terre selon différents lieux et moments de la journée.

La seconde pièce d'informations nécessaire pour effectuer une étude G est de déterminer les besoins du test en matière de mesure « relative » ou « absolue ». Une mesure relative est pertinente quand on cherche à situer des entités entre elles au sein d'une distribution ou d'une classification. Dans ce cas, la mesure consiste à déterminer la distance entre les entités (ex : un écart de cinq points entre deux élèves). Une mesure absolue détermine la position exacte de chaque entité sur une échelle de mesure (ex : un élève obtient une note de 15 sur 20). Son but est donc de définir la position d'un objet ou d'une personne sur une échelle donnée, indépendamment des autres individus. Même si ces deux mesures sont affectées par un certain niveau d'erreur de mesure, la mesure absolue y est la plus exposée.

Etude G : estimation des coefficients G et de l'erreur standard de mesure (SEM)

Le résultat d'une étude G est couramment résumé sous la forme d'un coefficient de généralisabilité (G), qui indique le degré pour lequel une procédure ou un instrument de mesure est capable de différencier avec fiabilité des individus ou des objets entre eux. En d'autres termes, le résultat nous informe sur la fiabilité à accorder à un dispositif, quels que soient les

éléments spécifiques qui le compose. Dans ce cas, on considère que les résultats ne dépendent pas (ou alors en proportion négligeable) des éléments particuliers et contingents qui ont composé les facettes au moment du test (les items du questionnaire, les juges, les utilisateurs, …). C'est pour cela que nous pouvons dire que les résultats sont généralisables à la population (finie ou infinie) de tous les éléments qui ont été utilisés pour développer l'instrument de mesure (l'univers des conditions d'observation admissibles). Cette conclusion est considérée comme raisonnable à partir d'un coefficient G égal au moins à 0.80. Il est calculé à partir d'un ratio entre la variance de différentiation et la variance totale :

$$\text{Coef. G} = \frac{\widehat{\sigma}_D^2}{\widehat{\sigma}_D^2 + \widehat{\sigma}_G^{2\prime}}$$

Où $\widehat{\sigma}_D^2$ et $\widehat{\sigma}_G^2$ représentent respectivement la variance estimée de différentiation et la variance estimée de généralisation (dont l'estimation varie en fonction du plan d'observation, d'échantillonnage, et de mesure). En fonction du type de différenciation souhaité, trois coefficients peuvent être calculés : un coefficient de généralisabilité relatif, un coefficient de généralisabilité absolu et un coefficient de généralisabilité critériel.

Le coefficient de généralisabilité relatif fut le premier à être formalisé. Il fut défini spécifiquement comme le coefficient de généralisabilité « G » par Cronbach et al. (1972) et est symbolisé par $E\hat{\rho}^2$. Il permet d'évaluer l'aptitude d'un dispositif instrumental à différencier de manière fiable des individus ou des objets en fonction de leurs positions relatives au sein d'une distribution de résultat. Dans le cas d'un plan croisé simple A x B, il correspond au coefficient Alpha de Cronbach. Le coefficient de généralisabilité absolu, ou coefficient de « *dépendabilité* » (Kane & Brennan, 1977), est symbolisé par φ. Il permet d'évaluer l'aptitude d'un dispositif instrumental à différencier de manière fiable des individus ou des objets en fonction des résultats (des scores) qui leur sont attribués au moyen d'une échelle de mesure absolue. Même si la formule ci-dessus est identique pour le calcul de ces deux coefficients, Φ est généralement moins élevé que $E\hat{\rho}^2$, car plus de composants contribuent à sa variance de généralisation. Enfin, le coefficient de généralisabilité critériel Φ(λ), permet de mesurer la fiabilité d'un test à situer des individus en fonction d'un certain seuil de référence (Kane & Brennan, 1977). Par exemple, pour un test cherchant à discriminer des personnes ayant eu la moyenne (10) sur un test en 20 points, Φ(10) indiquera à quel point ce protocole de mesure sera fiable pour situer les individus au-dessus ou en dessous de ce critère de score.

À côté de ces coefficients, il est possible également de calculer l'erreur standard de mesure (« *Standard Error of Measurement* » ou SEM), qui est une information importante à connaitre quand on évalue les qualités de mesure d'un instrument ou d'une procédure (Cronbach & Shavelson, 2004). Elle correspond à la racine carrée de la variance d'erreur et donne des informations directement interprétables sur l'imprécision d'un score de mesure. Par exemple, un score de 6 sur un test de 0 à 10 et avec une erreur standard de mesure de 0,5, sera compris dans un intervalle de 5,5 à 6,5 dans 68% des cas.

Etude D : étude de décision et d'optimisation (étude D)

L'étude G nous permet d'estimer la fiabilité d'un protocole en fonction des caractéristiques des données de l'étude-pilote qui a été réalisée. Néanmoins, il est possible, par simulation statistique, de calculer l'effet d'un changement de paramètre sur la fiabilité du protocole à partir des données récoltées. Ces études « D » nous permettent ainsi d'examiner le changement des coefficients de généralisabilité en fonction de divers paramètres de test, et de déterminer quelles sont les conditions idéales pour que la mesure soit la plus fiable possible. Pour cela, les études D permettent de manipuler le nombre de niveaux dans une facette, d'éliminer une facette ou un niveau spécifique à l'intérieur d'elle, ou d'en changer la nature.

La procédure d'optimisation la plus fréquente est la modification du nombre de niveaux pour la généralisation d'une facette. Par exemple, l'on peut décider d'augmenter le nombre d'items d'un questionnaire ou de juges, suite à un coefficient G insuffisant pour une procédure de test donnée. En effet, l'augmentation du nombre d'éléments d'une facette a pour conséquence de réduire l'erreur de mesure et donc d'augmenter la fiabilité du test. Cette simulation est rendue possible par l'utilisation de calculs statistiques dérivés de la formule « *prophétique* » de Spearman-Brown (Brown, 1910; Spearman, 1910). Quand un protocole utilise plusieurs facettes d'instrumentation, une stratégie d'optimisation courante consiste à augmenter le nombre de niveaux pour la facette qui génère une grande partie de la variance d'erreur et de diminuer la contribution d'autres ayant moins d'impact sur la fiabilité de la mesure. Ainsi, le nombre d'observation total sera équivalent et n'augmentera donc pas les coûts de mise en œuvre de la procédure de test, tout en augmentant sa fiabilité.

Un autre contrôle réalisable dans une étude D, quand le nombre de niveaux d'une facette est peu important, est de vérifier si un de ces niveaux n'a pas une influence disproportionnée sur la mesure, en affectant sa stabilité ou sa précision. Par exemple, un item d'un questionnaire peut se montrer dans les faits inconsistant avec le reste du test ou interagir de manière inappropriée avec un niveau particulier d'une autre facette (l'âge, le sexe, etc.). Il est ainsi possible de supprimer les niveaux atypiques de certaines facettes. Néanmoins, cette stratégie doit toujours être appliquée avec précaution, car il peut réduire la validité de la procédure de mesure, en changeant la nature de son univers de généralisation. Elle doit donc être utilisée seulement quand des raisons théoriques et/ou techniques l'autorisent.

Une autre stratégie consiste à modifier la nature ou le nombre de facettes dans le protocole de test pour augmenter sa fiabilité. En effet, quand une facette aléatoire génère plus d'inconsistance que prévu, une solution consiste à la fixer, en acceptant la perte conséquente de son pouvoir de généralisation. Par exemple, une facette « item » peut être fixée, ce qui peut améliorer la fiabilité de la procédure de test, mais restreindra les prochaines utilisations du protocole de test à ce seul questionnaire. Enfin, il est possible également de supprimer une facette du test, ce qui consiste à la « cacher », en fixant son niveau à un seul élément. Par exemple, on peut se rendre compte qu'un test, dont la prétention était d'être généralisé quel que soit l'âge, génère une variabilité sur cette facette trop importante pour tenir dans un seul test et devra être décliné en plusieurs (une procédure de test par catégorie d'âge). Il doit être reconnu que ces stratégies ont des implications importantes pour la mesure, car ces ajustements, s'ils améliorent la fiabilité du test, réduisent l'importance de l'univers de généralisation. Un

compromis est donc à trouver lors de l'élaboration du protocole de mesure afin que sa portée ne soit pas trop spécifique et que sa fiabilité reste raisonnable.

Méthodologie

L'objectif de cette étude est d'analyser le protocole de mesure multifacette qui a été utilisé pour comparer la qualité des recommandations de l'algorithme d'activation par propagation sémantique monocentrique (« M*onocentric Semantic Spreading Activation Algorithm* » ou MSSA), développé par les Bell Labs, à son concurrent, l'algorithme sVSM, utilisé dans le système de recommandation MORE (« *MOvie REcommendation* » ; Mirizzi, Di Noia, Ragone, Ostuni, & Di Sciascio, 2012). Cette partie présente la procédure d'évaluation de ces deux algorithmes (seulement la partie mesurant la pertinence des algorithmes, pour des soucis de clarté de la démonstration), ainsi que le plan d'étude utilisé pour les analyses de généralisabilité.

Procédure

Les participants ont évalué les résultats des algorithmes sur une interface neutre, mis en place avec la solution d'enquête en ligne Limesurvey[85]. L'évaluation se base sur cinq listes de recommandations de films, composées par le top 20 des résultats des deux algorithmes, soit 200 recommandations maximum à évaluer par participant. Les listes ont été randomisées, rassemblées dans une liste unique et les doublons ont été supprimés. De ce fait, les participants ne pouvaient pas prendre connaissance de la provenance des résultats. Les cinq films utilisés pour générer la liste ont été tirés aléatoirement de la liste « *50 films à voir avant de mourir* »[86]. Elle a été sélectionnée pour sa diversité : « *chaque film a été choisi comme l'archétype d'un genre ou d'un style particulier* »[87]. Les films sélectionnés aléatoirement étaient : 2001 l'odyssée de l'espace, Erin Brockovich, Terminator 2: le jour du jugement, Princesse Mononoke et Fight club. Pour chaque résultat une question sur le niveau de similarité d'expérience cinématographique était posée, accompagnée d'une échelle de Likert en quatre points :

« *Avec le film "2001 l'odyssée de l'espace", je pense que je vais vivre une expérience cinématographique similaire qu'avec "la planète des singes*": »

[ ] Pas du tout d'accord [ ] Plutôt pas d'accord [ ] Plutôt d'accord [ ] Tout à fait d'accord

Chacun des films recommandés était accompagné de la photo de l'affiche, du titre, du lien vers la page Wikipédia, du résumé, de la catégorie, ainsi que des personnes les plus importantes ayant participé à la création du film.

Plan d'étude pour l'analyse de la généralisabilité

Un plan d'étude a été mis en place afin d'estimer la fiabilité du protocole du test selon la théorie de la généralisabilité. Il a consisté à identifier les facettes d'intérêts, la nature de leurs relations et leurs types d'échantillonnage (Table 9). Puis, un objet d'étude a été déterminé, ainsi que le type de mesure visé par le protocole de mesure développé.

[85] www.limesurvey.org/

[86] http://www.film4.com/special-features/top-lists/top-50-films-to-see-before-you-die

[87] http://en.wikipedia.org/wiki/50_Films_to_See_Before_You_Die

Tableau 9 – Univers d'échantillonnage et niveau des facettes pour l'étude G

Abréviation	Nom de la facette	Niveau	Univers
A	Algorithme	**2**	Infini
F	Film	**5**	Infini
R	Recommandation	**20**	Infini
J	Juge	**15**	Infini

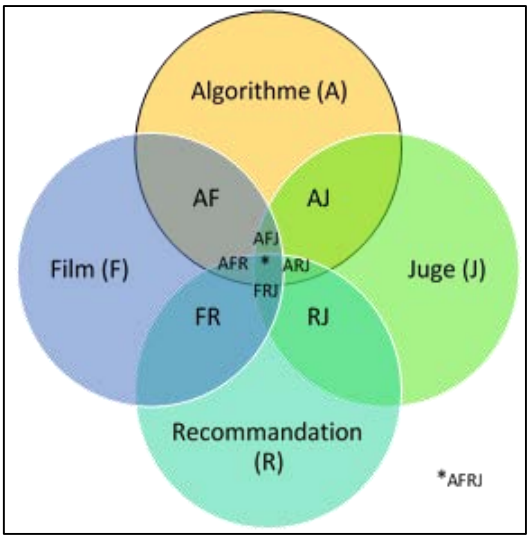


Figure 80 – Diagramme de partition de la variance pour le modèle AFRJ

Dans notre étude, les facettes « *Algorithme* », « *Recommandation* », « *Film* » et « *Juge* » ont été sélectionnées. La facette « *Algorithme* » représente l'ensemble des algorithmes de recommandations que nous cherchons à évaluer et à discriminer, en mettant au point ce protocole de mesure. La facette « *Recommandation* » contient les recommandations proposées par les différents algorithmes. La facette « *Film* » contient l'ensemble des œuvres cinématographiques pouvant être la cible de recommandations par les différents algorithmes. Nous avons pensé que c'était un point intéressant à prendre en compte lors de l'évaluation d'un algorithme, car nous pensions que leurs performances peuvent varier en fonction du type de film, sa notoriété, etc. Il a été donc décidé de limiter le champ d'étude au domaine du cinéma car, même si l'algorithme mis au point par les Bells Labs n'y est pas restreint, de nombreux autres y sont limités, et il n'est pas toujours possible, ni même sensé, de vouloir tester la performance d'un algorithme sur des domaines dont il n'a pas été conçu à la base. De plus, cela aurait alourdi considérablement le protocole de test. Néanmoins, il aurait été possible de créer une facette « Domaine », contenant de nombreuses catégories telles que la musique, le cinéma, les œuvres d'art, et plus encore, couplée à une autre facette « Eléments », qui aurait peuplé de différentes œuvres tous ces domaines. Enfin, la facette « Juge » contient les individus ayant eu pour mission de noter la pertinence des recommandations. Elle nous permet de quantifier la part de subjectivité individuelle contenue dans cette tâche d'évaluation.

Dans cette étude, toutes les facettes ont été croisées ensemble, selon un plan d'observation : A x F x R x J (Figure 80). Ainsi, tous les juges ont noté la pertinence de toutes recommandations, pour tous les films du test et pour tous les algorithmes. La variance totale (σ^2_T) pour ce test se décompose donc comme suit :

$$\sigma^2_T = \sigma^2_A + \sigma^2_F + \sigma^2_R + \sigma^2_J + \sigma^2_{AF} + \sigma^2_{AR} + \sigma^2_{AJ} + \sigma^2_{FR} + \sigma^2_{FJ} + \sigma^2_{RJ} + \sigma^2_{AFR} + \sigma^2_{AFJ} + \sigma^2_{ARJ} + \sigma^2_{FRJ} + \sigma^2_{AFRJ,e}$$

Les univers d'échantillonnage pour les quatre facettes ont été spécifiés comme infinis. Il s'agira donc de tester la fiabilité d'un protocole d'évaluation dont les éléments qui le composent (algorithmes, juges, recommandations, films) varient aléatoirement d'un test à l'autre. Leurs niveaux respectifs correspondant à l'étude G sont reportés dans le tableau 9. Comme le protocole d'évaluation vise à comparer divers algorithmes de recommandations entre eux, la facette « *Algorithme* » a été choisie comme facette de différentiation. Les facettes « Film », « Recommandation » et « Juge » ont été désignées comme facettes de mesure. Pour l'étude G,

les coefficients de généralisabilité relatifs et absolus ont été calculés, accompagnés de leurs erreurs standard de mesure. Des études D ont ensuite été menées, en faisant varier dans le protocole de test le nombre de film, de recommandation et de juge, pour en estimer les effets sur la fiabilité du test.

Résultats

L'enquête a été réalisée sur 15 personnes, dont 13 hommes et deux femmes. L'âge moyen était de 31,7 ans, et la plupart des participants étaient des chercheurs en informatique. Le nombre de films vus en moyenne par les participants était de 10,4 (σ = 8,66) par mois sur tout support.

Le tableau 10 expose l'analyse de la variance et sa décomposition en divers éléments. Il montre que la variation engendrée par l'interaction entre l'algorithme, la recommandation et le film, est particulièrement importante (25,8%), tout comme la variabilité résiduelle, non expliquée, (44,5%). En comparaison, la variation de différenciation n'est que de 1,9%. Cela est problématique pour un dispositif de mesure, car cela signifie que les écarts entre les éléments à discriminer sont petits par rapport à la fluctuation aléatoire générée par les facettes de mesure.

Tableau 10 – ANOVA et calcul des composants de la variance

Source de variations	Somme des carrés	ddl	Carré moyen	Composants				
				Aléatoire	Mixte	Corrigé	%	Erreur standard
A	41.301	1	41.301	0.016	0.016	0.016	1.9%	0.023
F	167.355	4	41.839	0.040	0.040	0.040	4.6%	0.043
R	242.155	19	12.745	0.063	0.063	0.063	7.3%	0.028
J	181.615	14	12.972	0.056	0.056	0.056	6.4%	0.023
AF	67.639	4	16.910	0.044	0.044	0.044	5.0%	0.033
AR	71.312	19	3.753	0.000	0.000	0.000	0.0%	0.017
AJ	5.079	14	0.363	0.000	0.000	0.000	0.0%	0.001
FR	244.352	76	3.215	-0.018	-0.018	-0.018	0.0%	0.026
FJ	101.325	56	1.809	0.035	0.035	0.035	4.0%	0.009
RJ	104.225	266	0.392	0.002	0.002	0.002	0.3%	0.005
AFR	286.015	76	3.763	0.225	0.225	0.225	25.8%	0.040
AFJ	23.981	56	0.428	0.002	0.002	0.002	0.2%	0.004
ARJ	102.708	266	0.386	0.000	0.000	0.000	0.0%	0.007
FRJ	394.368	1064	0.371	-0.009	-0.009	-0.009	0.0%	0.012
AFRJ	412.965	1064	0.388	0.388	0.388	0.388	44.5%	0.017
Total	2446.395	2999					100%	

En se basant sur les données du tableau précédent, le tableau 11 présente les résultats pour l'étude G pour le plan A/FRJ. Le coefficient de généralisabilité absolu obtenu lors de cette étude est faible (Φ = .38), ce qui traduit un manque de fiabilité du dispositif à déterminer la position exacte de chaque algorithme sur l'échelle de mesure de la pertinence des recommandations (de 0 à 3). Cela s'explique par une variance de différenciation faible, couplée à une variance d'erreurs élevée, notamment de la part de la facette Film et de l'interaction Film x Algorithme, qui totalise 62,8% de la variance totale pour la mesure absolue. La facette Recommandation et Juge, contribuent respectivement à 11,9% et 14% de la variance de mesure. Cela correspond, pour la facette Juge, à la variabilité de sévérité/clémence différente d'un participant à l'autre.

Néanmoins, on contraste également que leur niveau de sévérité est stable tout au long de leur évaluation, car les facettes d'interactions AJ, FJ et RJ ne représentent que 1,7% de la variance totale.

D'un autre côté, le coefficient de généralisabilité relatif pour cette étude est plus élevé, sans être encore satisfaisant ($E\hat{p}^2$= 0.59). La facette d'interaction AF est celle qui contribue le plus à la variance d'erreur (77,5%). Elle correspond à la variabilité de la pertinence des algorithmes en fonction du film, montrant que ces derniers sont plus performants pour certains films que pour d'autres. La facette AFR contribue également de manière importante avec 20% de la variance

Tableau 11 – Répartition de la variance et calcul des coefficients de généralisabilité

Variance de différenciation		**Variance d'erreur**				
Source	**Variance**	**Source**	**Variance d'erreur relative**	**%**	**Variance d'erreur Absolue**	**%**
A	0.01631		.....		.....	
	.....	F	.....		0.008	30.1%
	.....	R	.....		0.003	11.9%
	.....	J	.....		0.004	14.0%
	.....	AF	0.009	77.5%	0.009	32.7%
	.....	AR	(0.00)	0.0%	(0.000)	0.0%
	.....	AJ	(0.00)	0.0%	(0.000)	0.0%
	.....	FR	.....		(0.000)	0.0%
	.....	FJ	.....		0.0005	1.7%
	.....	RJ	.....		0.00001	0.0%
	.....	AFR	0.002	20.0%	0.002	8.4%
	.....	AFJ	0.00003	0.2%	0.00003	0.1%
	.....	ARJ	(0.00)	0.0%	(0.000)	0.0%
	.....	FRJ	.....		(0.000)	0.0%
	.....	AFRJ	0.0003	2.3%	0.0003	1.0%
Somme de la variance	0.016		0.011	100%	0.027	100%
Écart-type	0.128		**Écart-type relatif :**	0.106	**Écart-type absolu :**	0.163
Ep^2	0.59					
Φ	0.38					

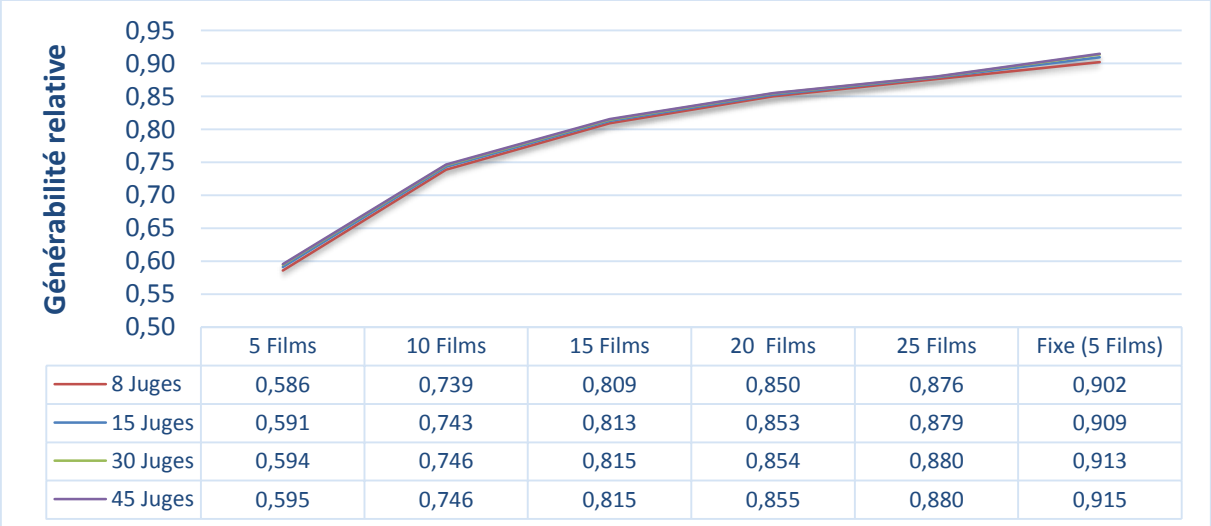


Figure 81 - Etude D sur l'évolution du coefficient de généralisabilité relatif en fonction du nombre de film et de juge (facette Recommandations composée de 15 éléments pur chaque algorithme testé)

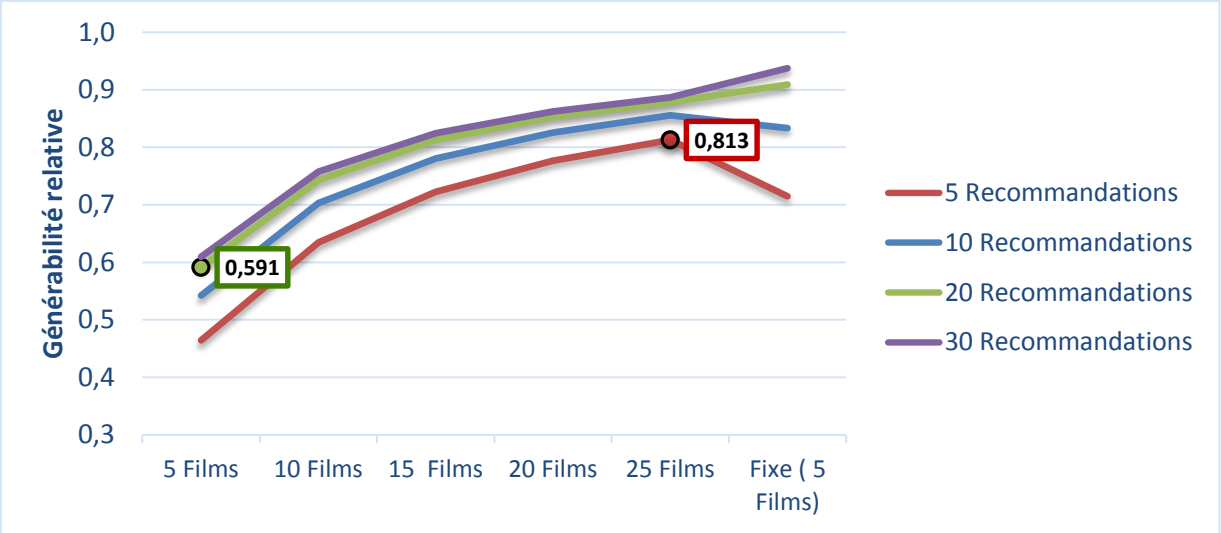


Figure 82 - Etude D sur l'évolution du coefficient de généralisabilité relatif en fonction du nombre de film et de recommandation (facette Juge composée de 15 éléments)

d'erreur relative. Enfin, on constate que les facettes Juges contribuent très faiblement à la variance de mesure, montrant que ces derniers maintiennent une évaluation consistante tout au long du test.

Les études D menées sur la variation du coefficient de généralisabilité relatif confirme les observations sur la contribution de la facette Film et Juge sur la fiabilité de la mesure (Figure 82). On constate ainsi un effet négligeable du nombre de juges sur le coefficient de généralisabilité relatif : pour 5 Films et 15 recommandations par algorithme testé, $E\hat{p}^2 = 0.586$ pour 8 juges et 0.595 pour 45 juges. Au contraire, le nombre de films testés à un impact important sur la qualité de la mesure : pour 15 juges et 15 recommandations par algorithme testé, $E\hat{p}^2 = 0.591$ pour 5 films et 0.879 pour 25 films. On constate également que fixer la facette Film (c'est-à-dire en faisant le choix d'utiliser la liste de films utilisée pour ce test pour tous les tests suivants) fait augmenter considérablement le coefficient de généralisabilité relatif ($E\hat{p}^2 =$ 0,902, pour 8 juges et 15 recommandations par algorithme testé). Cette différence d'effet sur la fiabilité de la mesure entre la facette Recommandation et Film, permet augmenter la fiabilité du test de 0.591 à 0.813, tout en n'augmentant que très peu la longueur du test, en passant simplement le nombre de recommandations de 20 à 5 et le nombre de films de 5 à 25.

Conclusion

Plusieurs enseignements peuvent être tirés de cette étude multifacette dont la portée peut être étendue aux domaines de l'utilisabilité et de l'UX. Premièrement, grâce à la théorie de la généralisabilité, nous avons vu qu'il est possible d'estimer précisément le nombre d'éléments nécessaires dans chaque facette pour que le test soit jugé comme suffisamment fiable, en se basant sur leurs instabilités respectives. Ainsi, en ayant constaté que la qualité des recommandations en fonction des films choisis varie en quantités importantes, nous avons pu remarquer qu'il valait mieux augmenter le nombre de films évalués, quitte à diminuer le nombre de recommandations à tester pour chacun de ces films. De plus, nous avons vu que le nombre de participants, pourtant faibles (n = 15), n'était pas un handicap pour le test, car leurs effets

sur la fiabilité de la mesure étaient presque négligeable en comparaison aux autres. Cette connaissance fine des rouages de la mesure permet d'ajuster un protocole aux plus près de ses besoins et de ses moyens. Par exemple, dans le cas où deux facettes d'instrumentation contribueraient à la même hauteur à générer de la variance d'erreur, il est possible de contrebalancer la réduction du nombre d'items d'une facette par l'augmentation du nombre d'items de l'autre, en permettant ainsi de réduire la contribution la plus coûteuse.

Deuxièmement, nous avons vu que, par rapport à la théorie classique de la mesure, la théorie de la généralisabilité permet de s'interroger sur des entités autres que des individus, grâce au principe de symétrie. Dans cette étude, ce sont les algorithmes de recommandation que nous avons cherché à "mesurer" (= différencier). Les participants, au contraire, ont constitué ici une facette instrumentale, c'est-à-dire considérés comme partie prenante d'un protocole de mesure. Cette propriété est particulièrement adaptée au domaine des IHMs, car il s'agit d'évaluer les systèmes numériques et non les individus. Nous pouvons ainsi étudier précisément la facette utilisateur (cachée dans les protocoles psychométriques classiques), qui, à cause de la subjectivité lors d'une interaction, pourrait contribuer grandement à la variabilité de l'évaluation, surtout lorsque l'on utilise des mesures de type UX.

Enfin, la théorie de la généralisabilité de la mesure nous offre un autre outil conceptuel qui peut nous être utile pour la construction de savoir méthodologique dans le domaine des IHMs : la portée d'un protocole de mesure. En effet, nous avons vu qu'une mesure peut avoir une portée relative ou absolue. Dans le cas d'usage présent, nous avons vu que seule l'utilisation de la mesure dans un cadre relatif était souhaitable. Le coefficient de généralisabilité relative est semblable à celui utilisé dans la théorie classique de la mesure, c'est-à-dire non généralisable à l'échantillon de l'étude. Il permet de comparer avec fiabilité des entités à l'intérieur du protocole de l'étude seulement. Ce type de mesure est efficace dans la plupart des situations expérimentales, c'est-à-dire là où toutes les conditions que l'on cherche à évaluer (=discriminer) se situent à l'intérieur d'un même protocole de test. Or, l'évaluation d'une application gagnerait en souplesse si l'on n'exige pas d'un protocole de tester toutes les autres applications concurrentes (ou futures…), avec les mêmes utilisateurs, voire les même items de questionnaire. Dans ce cadre, l'utilisation d'un coefficient de généralisabilité absolu, qui traduit l'habileté d'un test à pouvoir noter avec fiabilité une entité sur une échelle donnée, est une pratique à développer.

Plusieurs limites de l'étude actuelle sont également intéressantes à relever, pour relativiser les résultats obtenus, mais également pour en tenir compte dans les études à venir. La première limite concerne le nombre d'algorithmes testés et leur mode de désignation. Idéalement, les éléments de chacune des facettes devraient être tirés au sort et en nombre suffisant pour constituer des échantillons les plus représentatifs possible. Ici, les deux algorithmes qui constituent les éléments modèles pour représenter la facette de différenciation ont été sélectionnés selon un processus non aléatoire. La limite d'une telle procédure est que les écarts constatés entre ces deux algorithmes ne sont peut-être pas représentatifs des écarts moyens existant entre deux algorithmes tirés au sort. De ce fait, si l'écart moyen est plus important, le coefficient de généralisabilité du protocole de mesure est sous-estimé, et, s'il est plus important, il est surestimé. Dans les études à venir, il conviendra donc, dans la mesure du possible, de disposer de plus d'éléments pour représenter la facette de différenciation, et tiré au hasard, de

préférence. Néanmoins, cela met à jour une limite pratique importante des études utilisant la théorie de la généralisabilité : le nombre d'éléments à tester pour les participants. En effet, la plupart des études de généralisabilité utilise un plan d'étude entièrement croisé, ce qui permet d'étudier précisément l'effet de chaque facette et leurs interactions sur la fiabilité des mesures. Cependant, ce plan oblige les participants de tester toutes les combinaisons existantes, ce qui dans les faits peut poser des problèmes. Par exemple, dans cette étude, un juge doit évaluer vingt recommandations par film et par algorithme, soit 200 recommandations en tout. Si l'on multiplie le nombre d'éléments pour chacune des facettes, la passation du test peut devenir rapidement trop longue, dégradant sa réalisation. Un compromis doit donc être trouvé, entre un nombre suffisant par facette pour obtenir une estimation fiable des indicateurs, et un nombre mais pas trop élevé pour ne pas dégrader la réalisation de la passation.

Dans les études à venir, l'ensemble des outils conceptuels avancés par la théorie de la généralisabilité sera utilisé. Il sera particulièrement intéressant de voir la contribution de facettes rarement interrogées, comme celles des utilisateurs, et de voir quel minima s'impose dans des protocoles cherchant à mesurer des construits portant à l'utilisabilité ou à l'UX.

ETUDE 2 – MESURE MULTIMODALE ET A DISTANCE DE L'UTILISABILITE APPLIQUEE AUX SITES UNIVERSITAIRES

Introduction

L'évaluation de l'utilisabilité constitue encore une part importante de la pratique en ergonomie et se base en majeure partie sur la méthode des tests utilisateurs. Ces derniers peuvent être soit formatifs, soit sommatifs (Lewis, 2012). Les tests formatifs, plus informels et qualitatifs, permettent de détecter les problèmes affectant l'utilisabilité d'un produit afin de les corriger tout au long de son processus de développement. Les tests sommatifs, plus formels et quantitatifs, ont pour objectif de mesurer le niveau d'utilisabilité d'un système à un moment donné de son développement, en général plus avancé. Ils permettent ainsi la prise de décisions cruciales, comme celle du lancement d'un produit, si son niveau de maturité est considéré comme suffisant. C'est pourquoi les tests sommatifs doivent disposer d'une fiabilité suffisante pour appuyer de telles décisions.

La littérature dans le domaine des IHMs offre cependant peu d'aide sur la manière de mesurer l'utilisabilité et, en particulier, pour choisir les mesures qui lui sont adaptées (Hornbæk & Law, 2007). Une revue de ces mesures dans la recherche en IHM en a recensé plus de 54 types (Hornbak, 2006). Cette diversité a poussé de nombreux chercheurs à étudier les liens existant entre ces mesures (Fr0kjær, Hertzum, & Hornbæk, 2000; Hornbæk & Law, 2007; Nielsen & Levy, 1994) et à concevoir des moyens de les combiner (Jeff Sauro & Kindlund, 2005). Ces mesures composites, agrégeant des informations de différentes modalités, n'ont cependant pas ou peu été évaluées sous l'angle de leur qualité psychométrique, et il est donc difficile de les recommander, autrement que pour leur capacité à faciliter la communication. Or, leur utilisation devient de plus en plus incontournable, si l'on prend en compte l'avènement du Big Data et du Growth Hacking, basé sur le recueil à distance de données de toutes sortes (écrites, comportementales, occulométriques, physiologiques, …) et le traitement automatisé d'agrégations qui en résultent (de Vasconcelos & Baldochi, 2012; González, Lorés, & Granollers, 2008; Ivory & Hearst, 2001; Pour & Calvo, 2011; Ryu, 2009).

De manière plus générale, on peut constater que l'évaluation psychométrique des mesures de l'utilisabilité, discipline encore jeune, ne s'est pas encore adaptée aux caractéristiques propres au domaine. En effet, la majorité des approches statistiques utilisées, issues en grande partie des théories classiques de la mesure, ne prend pas suffisamment en compte certains paramètres agissant sur la fiabilité de la mesure dans le contexte de l'évaluation du numérique. Ainsi, ces évaluations psychométriques se concentrent principalement sur la consistance interne des outils de mesure, aux dépens de variabilité des utilisateurs et des interfaces, qui agissent pourtant tout autant sur la fiabilité des protocoles de mesure. Ces études, partielles, estiment ainsi la fiabilité d'une évaluation qu'au travers d'une seule de ses facettes, réduite généralement, elle aussi, qu'à la sphère d'un seul outil : les questionnaires, plus ou moins standardisés.

Ainsi, peu d'études ont été menées pour comparer, à la fois, la validité des mesures objectives, subjectives et leurs composés, tout en prenant en compte les facteurs principaux affectant la fiabilité d'une évaluation de l'utilisabilité. Une enquête portant sur la qualité de ces mesures, selon une approche de validation adaptée, permettrait d'estimer finement leur pertinence et d'avancer des moyens afin de mieux les articuler et les combiner. Cette étude vise à prendre en compte ces enjeux en testant la pertinence des modèles d'évaluation multimodaux, c'est-à-dire des protocoles de mesure qui s'appuient à la fois sur des données objectives et subjectives, à partir d'indicateurs composites. Elle se base sur une évaluation complète de ses principales facettes (outils de mesures, utilisateurs, interfaces, …), conformément aux spécificités du domaine des IHMs. Elle s'appuie sur des données issues d'une évaluation de l'utilisabilité de sites universitaires, effectuée à partir d'un logiciel de test utilisateur à distance (Evalyzer). Elle utilise un protocole de validations multifacettes, basé sur la théorie de la généralisabilité. Elle cherche ainsi à démontrer l'intérêt des indicateurs composites basés sur la récupération automatique de données par un système informatique.

L'évaluation de l'utilisabilité

L'évaluation de l'utilisabilité des portails web universitaires est un domaine typique faisant régulièrement usage de la méthode des tests évaluateurs pour mesurer la qualité des interfaces conçues. Ces derniers font souvent appel à quelques dizaines d'utilisateurs, devant réaliser un certain nombre de tâches représentatives sur l'interface en question. Des indicateurs de réussite et/ou de performance sont ainsi collectés, accompagnés parfois d'un questionnaire final de « satisfaction ». Ces différentes sources de données sont ensuite discutées individuellement, parfois confrontées, mais rarement combinées. Cette approche est illustrée par l'étude de Chaparro (2008). Pour évaluer l'utilisabilité d'un portail web d'une université du Kansas, il mit en place un test utilisateur composé d'un certain nombre de tâches à réaliser. Des données objectives (réussite et temps pour réaliser les tâches) et subjectives (évaluation de la difficulté des tâches et de la satisfaction générale) furent collectées pour chacun des participants. Ces données furent ensuite utilisées qualitativement, sans s'appuyer sur une méthode bien explicite, pour illustrer les différents problèmes d'utilisabilité constatés. L'étude de Roy, Pattnaik & Mall (2014) emprunte une méthodologie similaire : elle s'appuie, tour à tour sur des données collectées d'ordre objectif (réussite et temps pour réaliser les tâches, nombre de clics) et subjectif (questionnaire d'utilisabilité ASQ et WAMMI) pour tirer des conclusions sur la facilité d'usage de trois sites académiques.

Une approche différente, plus rare, vise à combiner ces sources de données pour se doter d'une mesure fiable et synthétique de l'utilisabilité. L'étude de Daher et Elkabani (2012), portant sur la mesure de l'utilisabilité des portails web pour six universités libanaises, en est un bon exemple. Si les données collectées, de nature objective (réussite et temps pour réaliser les tâches, nombre de clics et de changement de page) et subjective (questionnaires ASQ) sont similaires sur bien des points avec celles des études présentées précédemment, la façon de les exploiter est différente : elles sont standardisées et combinées pour calculer un score, dans l'objectif de synthétiser le niveau d'utilisabilité de chacun des portails web testés.

Ces deux approches illustrent le dilemme de fidélité-bande passante de Cronbach (1960), et son débat associé entre la préférence d'une évaluation riche ou d'une évaluation précise. Elle pose, dans le cadre de l'évaluation de l'utilisabilité, la question méthodologique suivante : « *Vaut-il mieux exploiter les mesures de l'utilisabilité de manière séparée ou est-il préférable de les combiner ?* ».

Les mesures d'utilisabilité : utilisation séparée ou combinée?

L'utilisabilité est une notion importante dans le domaine de l'ergonomie des IHMs. On compte, parmi ses nombreuses définitions, « *la capacité d'un système à être utilisé facilement et avec efficacité* » (Shackel, 2009b), « *l'efficacité, l'efficience et la satisfaction avec laquelle un utilisateur peut atteindre ses buts dans un environnement donné* » (ISO, 1998) ou encore, dans un sens plus large, la « *qualité dans d'utilisation* » d'un produit interactif (Bevan, 1999). Ces définitions, englobant de vastes domaines de l'interaction, expliquent en partie le nombre et la diversité des mesures de d'utilisabilité, qui tente d'opérationnaliser ce concept en fonction de ses nombreuses significations et sous-attributs associés (pour une revue, voir Alonso-Ríos et al.; 2009). Il n'est pas étonnant alors de voir apparaître des modèles de l'utilisabilité à dix dimensions et 26 sous-facteurs, qui sont eux-mêmes décomposés en 127 métriques spécifiques (Seffah, Donyaee, Kline, & Padda, 2006).

Le choix des mesures et de leurs utilisations combinées est également dirigé par la capacité à évaluer l'utilisabilité avec validité. En effet, l'utilisabilité est un « *construit* », au sens que la psychométrie lui donne, c'est-à-dire d'un concept non accessible directement par la mesure (comme l'intelligence) mais qui peut être approché par des moyens d'estimations plus ou moins valides. La difficulté d'obtention de mesures valides est très bien documentée dans la littérature sur l'évaluation des construits psychologiques depuis de nombreuses décennies (Cook & Campbell, 1979b). De même, la discussion sur la méthode à adopter pour mesurer la qualité des systèmes d'interactions numériques est un débat de fond qui ne date pas d'aujourd'hui. Elle a commencé à l'aube de l'ergonomie (Shackel, 1959), a été prolongée par la discussion autour de la « facilité d'utilisation » (Bennett, 1972), puis de l'utilisabilité (Shackel, 1981). Elle se poursuit encore de nos jours, notamment au travers du débat sur le choix des mesures lors de l'évaluation et de leurs pondérations. (Hornbak, 2006).

Le choix des mesures est conditionné en premier lieu par leurs qualités respectives en fonction du contexte d'évaluation. Par exemple, lors d'une évaluation formative, les données récoltées seront surtout d'ordre qualitatif, pour identifier les problèmes d'utilisabilité existant de manière exhaustive et en un temps court. Lors d'une évaluation sommative, les données récoltées seront avant tout quantitatives, pour permettre la mesure et la comparaison de différents produits sur une base solide, voire statistique. D'autres distinctions de mesure existent et sont sujettes à des discussions intensives sur leurs utilisations alternatives ou combinées. Il est courant ainsi, de faire la distinction entre les mesures objectives et subjectives (Meister, 1985; Yeh & Wickens, 1988), bien que cela résulte d'abord d'une séparation pratique plutôt qu'épistémologique (Muckler & Seven, 1992). Les mesures objectives de l'utilisabilité concernent les aspects que

l'on peut observer, telle que l'efficacité à réaliser une tâche. Les mesures subjectives de l'utilisabilité concernent les perceptions et les attitudes des utilisateurs envers une interface, une interaction ou un de ses résultats, via des méthodes de recueil le plus souvent verbales ou écrites. Une des raisons les plus avancées pour justifier l'utilisation conjointe de ces deux types de mesure est d'avoir une vue plus juste de l'utilisabilité d'un système, en confrontant ces données ou en les combinant. Par exemple, Tractinsky et Meyer (2001) ont trouvé des différences significatives entre plusieurs interfaces de l'expérience subjective du temps d'interaction alors que le temps objectif était le même. Ces différences entre mesures objectives et subjectives ont été retrouvées dans d'autres domaines, tels que l'évaluation de la charge de travail (Yeh & Wickens, 1988), ou de la performance à la tâche (Bommer, Johnson, Rich, Podsakoff, & Mackenzie, 1995). La confrontation de ces mesures nous permet ainsi d'enrichir et de raffiner notre jugement. Ces différences ont également été exploitées pour mettre au point de nouvelles mesures, par combinaison de celles-ci. Par exemple, Czerwinski et al. (2001) proposent une nouvelle mesure de l'utilisabilité, qui se base sur le ratio du temps « objectif » et de la durée d'interaction perçue « subjectivement » par l'utilisateur.

Une autre distinction classique des mesures est liée à la sous-division de l'utilisabilité en attributs séparés. Habituellement, les mesures sont classées en trois groupes, correspondant au standard de la norme ISO 9241 sur l'utilisabilité (ISO, 1998) : l'efficacité, « *la précision et la complétude avec laquelle un utilisateur accomplit une tâche spécifiée* », l'efficience, « *les ressources nécessaires pour accomplir une tâche spécifiée* », et la satisfaction, « *l'absence d'inconfort et les attitudes positives de l'utilisateur envers le produit* ». L'efficacité et l'efficience sont généralement mesurées par des métriques objectives, tels que le temps de réalisation et le nombre d'erreurs, alors que la satisfaction est mesurée généralement par des métriques subjectives, tels que les questionnaires. Hornbæk et Law (2007) montrent dans une méta-analyse sur l'évaluation de l'utilisabilité que l'utilisation de ces trois types de mesures dans une même étude est une pratique courante. En effet, sur 73 études examinées, 36 (49%) utilisent des mesures issues des trois attributs ; 30 études (42%) combinent des mesures d'efficacité/efficience, d'efficience/satisfaction ou d'efficacité/satisfaction ; 7 études (9%) ne collectent des mesures de l'utilisabilité que d'un seul attribut. De même, dans une étude menée sur 180 papiers issus de journaux en IHM, Hornbæk (2006) constate que les mesures d'efficacité, d'efficience et de satisfaction ne sont absentes que, respectivement, dans 22%, 18% et 38% des papiers examinés. Il constate également que la mesure de chacune de ces dimensions de l'utilisabilité s'appuie sur des métriques diverses. Les métriques les plus utilisées pour l'efficacité sont la réussite à une tâche, la précision (qui comprend le taux d'erreur) et la qualité du résultat (qui comprend la compréhension). Pour l'efficience, les métriques les plus utilisées sont le temps (dont le temps de complétion) et les patterns d'utilisation (dont la fréquence d'action). Pour la satisfaction, il s'agit de l'évaluation utilisateur de sa préférence, satisfaction et sentiment, à propos d'un produit qui lui est présenté. Compte tenu des enjeux et de la diversité des pratiques, de nombreuses normes ont été élaborées. Ainsi, un format-type de rapport, le CIF (« *Common Industry Format* »), a été mis au point, pour standardiser et formaliser les tests utilisateurs sommatifs (ANSI, 2001; ISO, 2006). Il propose de contrôler l'utilisabilité d'un produit à partir d'un certain nombre de mesures objectives (taux de complétion, temps de réalisation, nombre d'erreurs ou de demande d'aide, ….) et subjectives

(questionnaire ASQ, SUMI, SUS, …) qui couvrent ainsi les trois dimensions de l'utilisabilité précédemment citées.

Ainsi, à côté du besoin de plus en plus fort d'utiliser plusieurs méthodes de recherche pour couvrir un sujet d'étude (Filippi & Barattin, 2012; Remus & Wiener, 2010; Wilson, 2006), un consensus fort se dégage actuellement sur l'utilisation de plusieurs métriques pour couvrir au mieux l'utilisabilité d'un produit. Les avantages avancés d'une telle démarche sont similaires : (i) la validation croisée des données, obtenue par différentes sources, augmente la fiabilité, la validité et la robustesse de l'estimation, et, de ce fait, la confiance que l'on peut avoir des résultats obtenus (Creswell, 2003; Wilson, 2006) ; (ii) les approches plurielles et multimodales permettent d'étudier un domaine sous toutes ses coutures, de le couvrir convenablement ou de l'étendre, et ainsi de le comprendre en profondeur ou encore de découvrir de nouveaux paradoxes qui stimuleront les recherches à venir (Kaplan & Duchon, 1988; Mingers, 2001). Néanmoins, une question méthodologique divise chercheur et praticien : « Vaut-il mieux exploiter ces mesures de manière séparée ou est-il préférable de les combiner ? ».

Dans les faits, la majorité des études qui font le choix d'utiliser différentes mesures de l'utilisabilité les exploite ensuite de façon séparée (Hornbak, 2006). Néanmoins, Hornbak (2006) constate également qu'un certain nombre d'études font le choix de combiner des mesures de l'utilisabilité en une seule mesure, en reporte la valeur combinée et procède à des tests statistiques sur cette combinaison. Des méthodes de combinaison de mesures ont été spécialement mises au point dans le paradigme de l'utilisabilité (Jeff Sauro & Kindlund, 2005; T. Tullis & Albert, 2008), telle que la méthode SUM. Ces méthodes se basent généralement sur des procédures de standardisation (tel que le z-score), de pondération et d'agrégation des mesures. Ainsi, Chadwick-Dias, McNulty et Tullis (2003) transforment, par la méthode des z-scores, puis agrègent, par une combinaison à poids égaux, les mesures du temps et de réussite à la tâche, afin de comparer l'utilisabilité de deux prototypes d'itérations successives. En étudiant des interfaces de boîte mail, Whittaker et al. (2002) constatent que les ressources déployées lors de la réalisation des tâches varient grandement d'un utilisateur à l'autre. Pour contrôler ce biais, ils mettent au point une mesure normalisée de la performance utilisateur avec la formule : « *Qualité de la solution* » / « *Temps nécessaire pour réaliser la solution* » (Whittaker et al., 2002, p. 279). Afin de mesurer précisément l'impact de l'âge des utilisateurs sur la facilité d'utilisation d'un site d'information de santé, Pak, Price & Thatcher (2009) décide de créer une variable composite de la performance, à partir de trois mesures : le temps de réalisation de la tâche, le nombre de clics et le nombre d'erreurs. Chacune de ces mesures est ainsi normalisée (par une transformation z-score), puis agrégée (par une pondération à poids égaux) en une seule mesure. Le bénéfice avancé par les auteurs de créer une variable composite est de disposer d'une mesure de plus grande stabilité. Une des études les plus abouties visant à mettre au point un score d'utilisabilité à partir de la combinaison de plusieurs métriques a été menée par Jeff Sauro et Erika Kindlund (2005). En se basant sur la méthode des six Sigma (Breyfogle, 1999), ces derniers mettent au point un score, le SUM (« Single Usability Metric »), composé de quatre mesures de l'utilisabilité : réussite, temps, erreurs à la tâche et évaluation subjective de la difficulté de la tâche (via le questionnaire ASQ). À partir de données issues de tests utilisateurs réalisés sur une période de deux ans, composés de 129 participants et sur 57 tâches prédéterminées, les auteurs effectuent une analyse en composante principale (ACP) pour

cerner la contribution de chacune de ces quatre mesures sur le score global d'utilisabilité. Ils trouvèrent que les quatre mesures contribuèrent chacune de manière significative et égale. De ce fait, ils décidèrent que le mode de calcul du SUM se baserait sur une agrégation à poids égaux de ces quatre mesures standardisées. L'étude de Daher et Elkabani (2012) utilise ce modèle pour comparer l'utilisabilité de six portails web universitaires.

Parmi les arguments avancés principalement pour justifier l'utilisation de mesures combinées, on note la simplification des données et l'augmentation de la validité de la mesure. En effet, dans un monde où les décisions se prennent de plus en plus rapidement, le besoin d'indicateurs synthétiques et faciles à comprendre commence à s'imposer. Sauro et Kinlund (2005) pointent ironiquement ce fait : les métriques de l'utilisabilité devraient être plus faciles à utiliser. La complexité d'analyse et de présentation des données rend l'utilisabilité difficile à digérer. En effet, l'analyste est mis au défi de présenter plusieurs mesures d'utilisabilité, qui doivent exprimer clairement les aspects utilisables et inutilisables d'un produit, sans surcharger mentalement les chefs d'entreprise ou de promouvoir, par inadvertance, une métrique sur une autre. Ainsi, pour augmenter la pertinence et l'influence stratégique des mesures de l'utilisabilité lors d'une communication, les analystes doivent être en mesure de présenter synthétiquement la facilité d'utilisation, en utilisant un seul score si possible, et sans en sacrifier la précision (Jeff Sauro & Kindlund, 2005; T. Tullis & Albert, 2008). De plus, la fusion de divers sources de données s'imposera d'autant plus que le nombre d'informations sur l'utilisabilité d'un produit ne va cesser d'augmenter (González et al., 2008). Le contre-argument classique face à ce courant est que la combinaison de mesures peut cacher des patterns sous-jacents dans les données (Zhai, 2004), et que l'on risque de perdre des informations importantes dans l'opération, surtout si les mesures utilisées sont peu liées (Hornbæk & Law, 2007). Or, si ce constat se tient dans le cadre d'une approche formative, c'est-à-dire quand l'on cherche à diagnostiquer en détails une situation pour l'améliorer, elle n'est pas défendable dans un cadre sommatif, où l'on cherchera à mesurer avec la plus grande fiabilité l'état d'une situation.

C'est bien cet argument qui est avancé pour justifier l'utilisation de mesures combinées, c'est-à-dire l'amélioration de la fiabilité et de la validité de mesures (Hulin et al., 2001; Pak et al., 2009; Sauro & Kindlund, 2005). Or ses adversaires dénoncent des biais dans la procédure de création et d'utilisation des mesures combinées qui risque au contraire de nuire à sa qualité. Ainsi, il est pointé que la validité d'un score d'utilisabilité, tel que celui proposé par Sauro et Kindlund (2005), est limité par les mesures qui sont inclues ou exclues de sa procédure d'agrégation (Fr0kjær et al., 2000). De plus, les mesures choisies et leurs pondérations respectives, peuvent être plus ou moins pertinentes en fonction de la situation à observer. Par exemple, dans un contexte militaire ou chirurgical, le nombre d'erreurs est un facteur de qualité du produit beaucoup plus critique que dans d'autres domaines. Enfin, il y a une difficulté à considérer comme possible la fusion de métriques peu corrélées entre-elles, indiquant des différences de nature trop marquées pour être synthétisées sous un indicateur unique (Hornbæk & Law, 2007). À l'inverse, les défenseurs des mesures combinées déclarent qu'elles tirent justement leur force de la quantité et de la diversité métrique utilisée pour leur élaboration. La quantité des mesures utilisées permet, selon un principe central dans la CTT (« Classical Test Theory » ; Novick, 1966), d'améliorer la stabilité de la mesure, en neutralisant une part des fluctuations aléatoires dans les données (Hulin et al., 2001). En plus, cela simplifie les analyses,

et réduit le risque d'inflation des erreurs de type I (Hornbak, 2006), qui augmentent en fonction du nombre de tests réalisés. La diversité des mesures augmente également la validité d'un score. En effet, nous pouvons faire face à deux erreurs : celle issue du bruit (erreur non systématique), bien décrites dans la théorie classique de la mesure et dans la théorie de l'information de Shannon (1948) ; et celle issue d'un point de vue biaisé (erreur systématique). Cette erreur de partialité nous rappelle qu'il faut essayer de voir toutes les facettes et tous les aspects d'une même réalité pour la saisir. Cela se réfère à la notion de complexité chez Edgar Morin : « *Quand je parle de complexité, je me réfère au sens latin élémentaire du mot "complexus", "ce qui est tissé ensemble". Les constituants sont différents, mais il faut voir comme dans une tapisserie la figure d'ensemble. Le vrai problème (de réforme de pensée) c'est que nous avons trop bien appris à séparer. Il vaut mieux apprendre à relier* » (Morin, 1995). Combiner des informations de sources diverses, nous permet donc d'appréhender une entité de manière plus complète. De plus, l'indépendance des sources nous permet de neutraliser certains biais inhérents à l'acquisition de certaines données et d'obtenir une mesure globale plus fiable. Notre système nerveux tire parti de ce principe en combinant quasi-optimalement des sources d'informations sensorielles multiples (Ernst & Banks, 2002). En s'appuyant sur le principe du maximum de vraisemblance, issu des lois de l'inférence Bayésienne, la perception se base sur une moyenne pondérée des valeurs suggérées par chaque indice. Les pondérations sont fonction de la fiabilité (« *reliability* ») des indices, soit l'inverse de la variance. La fiabilité totale est calculée par la somme des fiabilités, car l'information est additive pour des signaux indépendants : la fiabilité augmente donc lorsque l'on dispose de plus de modalités. En revenant à notre sujet, disposer de mesures fiables (réductions des erreurs aléatoires), diverses et indépendantes, augmenterait les chances de construire une mesure composite qui soit valide.

Or, les études actuelles, contradictoires, visent, tout au mieux, de tester le niveau de corrélation entre les différentes mesures de l'utilisabilité. Les premières études montrent que des mesures objectives, telles que le temps et le taux d'achèvement d'une tâche, ne se corrèlent pas nécessairement à des mesures subjectives, telles que la préférence des utilisateurs et la satisfaction perçue (Fr0kjær et al., 2000; Kissel, 1995; Nielsen & Levy, 1994). Les résultats de la méta-analyse menée par Hornbæk et Law (2007) montrent que l'efficacité, l'efficience et la satisfaction (à savoir les trois mesures d'utilisabilité prototypiques) sont corrélées à un niveau faible à moyen. Sauro et Lewis (2009), dans une autre méta-analyse, ont trouvé une relation significative entre des scores issus de questionnaires standardisés d'utilisabilité et d'autres mesures d'utilisabilité prototypiques, telle qu'une corrélation entre la complétion d'une tâche de 0,24 dans le contexte d'une évaluation post-test de la satisfaction après l'achèvement des tâches multiples, et une corrélation moyenne de 0,51 dans le cadre de mesure de la satisfaction immédiatement après l'exécution d'une tâche unique. Faisant suite à cette recherche, Kortum et Meres (2014) ont mené deux études dans lesquelles ils ont trouvé des corrélations significatives, bien que très variables, entre les scores d'efficacité du système et les scores d'utilisabilité perçus (Etude 1, $r = 0.21$; Etude 2, $r = .73$).

Néanmoins, contrairement à ce que l'on voudrait nous faire croire, l'existence de corrélation forte ou faible entre ces mesures ne justifie pas l'utilisation ou la non-utilisation de mesures combinées. Il s'agirait plutôt de prouver que ces mesures permettent de mesurer l'utilisabilité avec une plus grande fiabilité et validité. Ces tests, d'ordre psychométrique, peinent déjà à

s'imposer pour des mesures plus simples, et encore beaucoup plus rares dans le cas de mesures combinées.

L'évaluation psychométrique des mesures d'utilisabilité

L'ANSI, dans leur rapport standardisé CIF (Common Industry Format), recommande d'utiliser comme mesure de l'utilisabilité le taux de complétion de la tâche (pour une mesure de l'efficacité), le temps de réalisation de la tâche (pour une mesure de l'efficacité), et la moyenne des scores de satisfaction des participants (ANSI, 2001). Ces mesures sont, par ailleurs, utilisées couramment dans la pratique de l'évaluation de l'utilisabilité (Hornbæk & Law, 2007).

Il est intéressant de constater que, d'un point de vue psychométrique, l'estimation de la fiabilité des mesures objectives n'a que peu intéressé les chercheurs, qui donnent, tout au plus, quelques précautions d'utilisation. Par exemple, pour le taux de complétion ou l'occurrence d'un problème, on recommande de calculer des intervalles de confiance réalistes, telles qu'avec la méthode d'ajustement de Wald (Lewis & Sauro, 2006; Sauro & Lewis, 2005). Pour le temps de tâche, Sauro (2008) préconise d'utiliser des tests statistiques adaptés à la taille de l'échantillon de test, voire de se concentrer sur le temps d'amélioration relatif de la tâche entre deux itérations. De plus, il vaut mieux reporter la moyenne géométrique du temps de tâche que la médiane, qui est généralement biaisée positivement (Cordes, 1993; Sauro & Lewis, 2010).

D'un autre côté, on constate que de nombreux efforts de validation psychométriques se sont concentrés sur les questionnaires standardisés, qui ont profité des outils existants, hérités de la psychologie, et particulièrement adaptés à l'évaluation des questionnaires. Ils sont conçus pour un usage répété, sont généralement composés d'un ensemble spécifique de questions, présentées dans un ordre spécifié, à l'aide d'un format spécifié, et avec des règles spécifiques de la production de la mesure. Ces questionnaires d'utilisabilité standardisés ont été introduits dans le champ des IHMs dans les années 80-90 (Brooke, 1996; Chin et al., 1988; Kirakowski & Corbett, 1993; Lewis, 1995), notamment sous l'influence des psychologues expérimentaux qui rejoignirent le domaine (Lewis, Utesch, & Maher, 2015), et n'ont cessé de se développer depuis. Une tendance croissante consiste à favoriser l'utilisation des échelles courtes en raison de leur vitesse et leur facilité d'administration, soit comme des sondages en ligne pour les clients ou après un test d'utilisabilité. C'est ce qui a fait, entre autres, le succès de l'échelle en 10 items SUS (« *System Usability Scale* » ; Brooke, 1996), citée dans plus de 600 publications (Sauro, 2011). Ce mouvement s'est encore amplifié aujourd'hui, avec des échelles très petites, pouvant être réduites à seulement deux items (Finstad, 2010, 2013; Lewis et al., 2015). Les échelles d'utilisabilité se divisent en deux familles : celles utilisées à la fin d'une étude, comme le QUIS (« *Questionnaire for User Interaction Satisfaction* » ; Chin et al., 1988), le SUMI (« *Software Usability Measurement* Inventory » ; Kirakowski and Corbett, 1993) ou le PSUQ (« *Post-Study System Usability Questionnaire* »; Lewis, 2002) ; et celles utilisées après chaque tâche, comme le ASQ (« *After-Scenario Questionnaire* » ; Lewis, 1991), le SEQ (« *Single Ease Question* »; Tedesco & Tullis, 2006), ou le UME (« *Usability Magnitude Estimation* »; McGee, 2003). Sans surprise, ces questionnaires se sont montrés plus fiables que les questionnaires ad hoc (Hornbæk & Law, 2007; Hornbæk, 2010; Sauro & Lewis, 2009) , même si, dans la pratique, on constate qu'ils ne sont que marginalement utilisés. En effet, dans une méta-analyse de 2006, sur 112

études cherchant à mesurer la satisfaction des utilisateurs, on constate que seulement 12 études utilisent un questionnaire standardisé, et seulement 10 études cherchent à contrôler la fiabilité de la mesure, par des procédures comme l'alpha de cronbach (Hornbak, 2006) ; ce qui nous montre que la psychométrie n'est ni la discipline la plus appréciée, ni la plus courante dans le domaine des IHMs (Lindgaard & Kirakowski, 2013).

On évalue généralement la qualité psychométrique des questionnaires standardisés à partir de mesures de la fiabilité , de la validité et de la sensibilité (Nunnally & Bernstein, 1994). La fiabilité, comprise comme la régularité de la mesure, est calculée de plusieurs manières, telles que par la méthode du test-retest ou des formes parallèles (split-half). Néanmoins, la méthode la plus commune pour mesurer la fiabilité d'un test est celle du coefficient alpha (ou coefficient de cronbach), qui mesure la consistance interne d'un questionnaire (Cronbach, 1951). Cette méthode est très largement utilisée pour valider les échelles ou les sous-échelles des questionnaires standardisés de l'utilisabilité, tels que pour le QUIS (Chin et al., 1988), le SUMI (Kirakowski, 1996), le PSSUQ (Lewis, 2002), le SUS (Bangor, Kortum, & Miller, 2008; Lewis & Sauro, 2009), le WAMMI (Kirakowski & Dillon, 1998), ou encore le SUPR-Q (Sauro, 2011b). Le coefficient alpha est utilisé également pour de très petits questionnaires, tels que le UMUX et le UMUX-Lite (Finstad, 2010; Lewis et al., 2015). Il ne peut pas, par contre, être utilisé pour des échelles mono-items, telles que le SEQ (Tedesco & Tullis, 2006) ou l'UME (McGee, 2003), car il nécessite au moins deux items pour son calcul.

La validité d'un test peut être comprise comme le degré pour lequel un test proposé permet d'identifier le construit mesuré. D'innombrables méthodes existent pour mesurer la validité d'un test, comme la validité concurrente, prédictive, convergente, discriminante ou encore de construit. La validité convergente et prédictive d'un instrument désigne le degré de corrélation entre les indicateurs qu'il permet d'obtenir et ceux qui ressortent d'autres instruments communément admis et qui sont mesurés soit au même moment (validité concurrente), soit à une période ultérieure (validité prédictive). Dans le cadre de la validation des questionnaires d'utilisabilité, l'estimation de la validité concurrente est une pratique courante. Elle est effectuée en calculant la corrélation entre les questionnaires testés et des mesures de référence. Ces mesures peuvent être soit objectives, soit subjectives (souvent d'autres questionnaires validés), soit les deux à la fois. La mesure objective la plus utilisée pour contrôler la validité concurrente d'un questionnaire d'utilisabilité est le taux de réussite à une tâche (cf. PSSUQ, ASQ, SEQ et UME; Lewis, Henry, & Mack, 1990; Lewis, 1995; Sauro & Dumas, 2009). Le temps nécessaire et le nombre d'erreurs commises lors de la réalisation d'une tâche sont également souvent utilisés (cf. SEQ et UME ; Sauro & Dumas, 2009 ; McGee, 2003). Plusieurs de ces mesures peuvent être utilisées à la fois lors d'un même test, ou de manière composée (cf. l'utilisation de la combinaison des mesures du temps et taux de réussite à une tâche pour contrôler la validité du SEQ ; Tedesco & Tullis, 2006). Parmi les mesures subjectives utilisées pour contrôler la validité concurrente, c'est le questionnaire SUS qui fait le plus référence (cf. la validation du SUPR-Q, SEQ, UME, UMUX ou encore de l'UMUX-lite ; Sauro, 2011b ; Sauro & Dumas, 2009 ; Finstad, 2010 ; Sauro & Lewis, 2012 ; Lewis et al., 2015). Pour le contrôle de la validité prédictive, Davis (1989) fait partie des seuls à utiliser ce type de validation. En effet, dans le cadre du développement du modèle d'acceptation de la technologie

(« *Technology Acceptance Model* » ou TAM), il effectue la corrélation entre la facilité d'utilisation perçue (« *Perceived Ease Of Use* » ou PEOU) et l'usage futur de l'outil.

Le deuxième type de validité souvent contrôlé pour les questionnaires d'utilisabilité est celui de la validité de construit, mise à l'épreuve grâce à deux techniques : les analyses factorielles et les matrices multi-traits multi-méthodes (M.T.M.M.). L'analyse factorielle contrôle si les réponses données aux items d'un instrument se regroupent comme elles le devraient théoriquement. Par exemple, Wang et Senecal (2007) utilisèrent une analyse factorielle confirmatoire pour valider la structure en trois facteurs de leurs questionnaire d'utilisabilité perçu ; et Borsci et al. (2015) utilisèrent la Théorie de la Réponse à l'Item multidimensionnelle (« *Multidimensional Latent Class Item Response Theory models* » ; Bacci, Bartolucci, & Gnaldi, 2014) pour confirmer la structure unidimensionnelle du questionnaire SUS. Néanmoins, la plupart des méthodes d'analyses factorielles utilisées dans le domaine sont tirées de méthodes exploratoires, telle que l'analyse en composante principale (cf. la validation du QUIS, SUMI, PSSUQ, ASQ, SUS et UMUX ; Chin et al., 1988 ; Kirakowski, 1996 ; Lewis, 1991 , 2002 ; Borsci, Federici, & Lauriola, 2009 ; Lewis, Utesch, & Maher, 2013 ; Lewis et al., 2015). Ces méthodes peuvent aider le chercheur à réduire ou structurer les items d'un questionnaire à construire mais sont peu adaptées pour une validation de nature factorielle. Une autre méthode couramment utilisée pour déterminer la validité de construit est la matrice multi-traits multiméthodes (MTMM) développée par Campbell et Fisk (1959). Elle permet de comparer les résultats obtenus via l'utilisation de différentes méthodes sur un même trait (construit), et ceux de différents construits à l'aide d'une même méthode. Ainsi, il devrait y avoir une forte corrélation entre les items ou instruments de mesure qui reposent sur des construits théoriquement similaires (preuve d'une validité convergente) ; et des corrélations faibles entre des instruments mesurant des construits théoriquement différents (preuve d'une validité discriminante). Cette procédure a été utilisée pour contrôler la validité de construit du modèle d'acceptation de la technologie de Davis (1989) et du questionnaire de mesure de la qualité Web d'Aladwani et Palvia (2002).

En corolaire, si un questionnaire est fiable et valide, il devrait être également sensible aux différentes variations d'états du construit mesuré. La sensibilité (ou discrimination) d'un test peut être comprise comme la capacité d'un instrument de mesure à différencier efficacement entre différents états d'intérêt. A l'inverse, un bon test doit être insensible aux variables extérieures qui pourraient parasiter l'estimation exacte du construit mesuré. Il n'y a pas de mesure directe de la sensibilité similaire à celles existantes pour la fiabilité ou la validité. Une mesure indirecte de la sensibilité consiste à comparer différentes situations et de voir si le questionnaire permet de les différencier statistiquement. La taille de l'effet (mesure de la force de l'effet observé d'une variable) et la taille minimum de l'échantillon nécessaire pour arriver à la signification statistique sont également utilisées pour mesurer la sensibilité d'un test (Sauro & Lewis, 2012). En raison de l'objectif même visé par les questionnaires d'utilisabilité, le facteur manipulé couramment pour tester sa sensibilité est représenté par différents sites web, produits ou systèmes, de qualité différente, et dont le questionnaire se doit de pouvoir discriminer (cf. la validation du PSSUQ, SUS, ASQ, UME, QUIS, CSUQ, SMEQ, SUMI, SUPR-Q, et l'UMUX ; Lewis, 2002 ; Bangor et al., 2008 ; Lewis & Sauro, 2009 ; Lewis et al., 1990 ; Tullis & Stetson, 2004 ; Sauro & Dumas, 2009 ; Kirakowski, 1996 ; Chin et al.,1988 ;

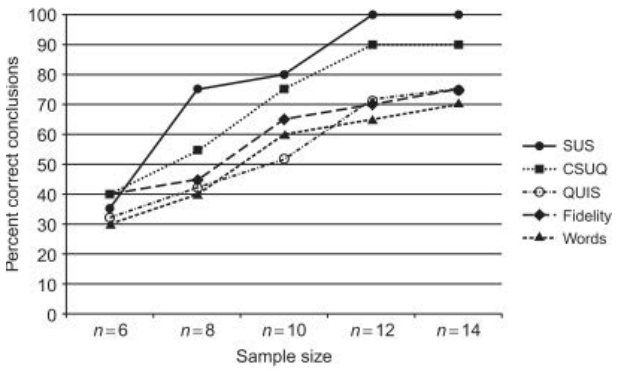


Figure 83 – Sensibilité relative de cinq méthodes d'évaluation de l'utilisabilité *(Tullis & Stetson, 2004)*

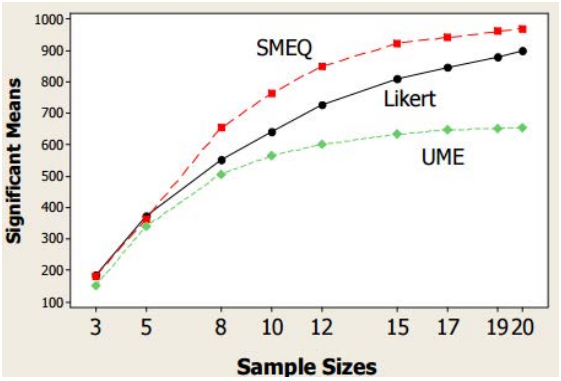


Figure 84 – Sensibilité relative de trois types de questions mesurant l'utilisabilité *(Sauro & Dumas, 2009)*

Sauro, 2015 ; Finstad, 2010). La sensibilité des questionnaire a également été testée en utilisant différentes tâches et scénario sur un même système (cf. la validation de l'ASQ et l'UME ; Sauro & Dumas, 2009; Tedesco & Tullis, 2006), voire sur la capacité du test à discriminer certaines caractéristiques utilisateurs pouvant influencer l'utilisabilité perçue, telle que l'expertise (cf. la validation du PSSUQ, SUS et UMUX-Lite ; Lewis et al., 1990; Lewis et al., 2015). Une méthode plus précise consiste à tester la sensibilité d'une ou plusieurs mesures de l'utilisabilité, en faisant varier la taille de l'échantillon. Pour cela, les auteurs sélectionnent aléatoirement un grand nombre de sous-échantillons de données, selon différentes tailles, et calculent pour chacun le nombre de tests statistiques significatifs (Sauro & Dumas, 2009; Tedesco & Tullis, 2006; Tullis & Stetson, 2004). Par exemple, en appliquant cette méthode d'échantillonnage aléatoire en fonction de différentes tailles d'échantillons, Tullis & Stetson (2004), comparent la sensibilité de cinq méthodes d'évaluation de l'utilisabilité (Figure 83) et Sauro & Dumas (2009) de trois types de questions (Figure 84). Enfin, certains auteurs cherchent, au contraire, à montrer l'insensibilité (dit autrement, la résistance) du questionnaire à certains facteurs, comme son support (cf. format papier ou en ligne du QUIS ; Slaughter, Harper, & Norman, 1994), sa complétude (cf. PSSUQ ; Lewis, 2002), l'adaptation de certains items au type de système mesuré (Lewis & Sauro, 2009), ou encore le sexe des utilisateurs qui le complètent (cf. PSSUQ ; Lewis, 2002).

Les limites des approches psychométriques actuelles

Nous constatons de nombreuses limites aux études existantes devant valider la qualité des mesures de l'utilisabilité développée. En premier lieu, elles sont très limitées quand elles concernent des mesures objectives, telles que le temps de tâche ou le taux d'erreur ; et sont quasi-inexistantes quand il s'agit de mesures plus élaborées encore, telles que des combinaisons de plusieurs mesures objectives et/ou subjectives. Or, ces mesures sont vues comme des sources fortes et complémentaires pour améliorer l'évaluation de l'utilisabilité d'un produit (Bangor et al., 2008) , et il convient donc de mieux l'appréhender. En deuxième lieu, on constate que, là même où les données psychométriques sont les plus abondantes, c'est-à-dire dans le cadre de validation de questionnaires subjectifs, elles sont parcellaires et peu adaptées aux IHMs. Ces lacunes s'expliquent très bien par l'histoire du domaine, encore jeune et peu adapté encore aux caractéristiques propres des IHMs. En effet, la majorité des approches statistiques utilisées,

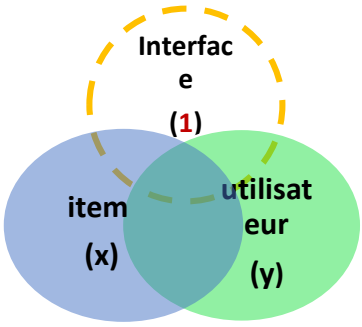


Figure 85 – Illustration en cercles pleins des facteurs pris en compte dans les modèles psychométriques actuels

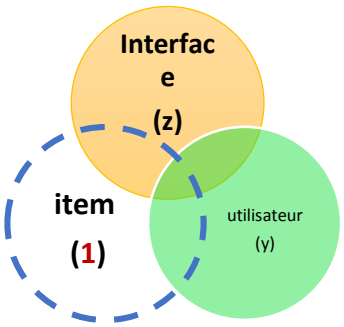


Figure 86 – Illustration en cercles pleins des facteurs pris en compte dans le calcul de la fiabilité inter-juges

issue en grande partie des théories classiques de la mesure, ne prend pas suffisamment en compte certains paramètres agissant sur la fiabilité de la mesure dans le contexte de l'évaluation du numérique. Ils se basent sur des modèles psychométrique bi-facettes, peu adaptés au domaine des IHMs, comportant, au minimum, trois facettes (Figure 85). Ainsi, ces évaluations psychométriques se concentrent principalement sur la consistance interne des outils de mesure, aux dépens de variabilité des utilisateurs et des interfaces, qui agissent pourtant tout autant sur la fiabilité des protocoles de mesure. Ces études, partielles, estiment ainsi la fiabilité d'une évaluation qu'au travers d'une seule de ses facettes, réduite généralement, elle aussi, qu'à la sphère d'un seul outil : les questionnaires, plus ou moins standardisés. Or, ces deux facettes livrent des informations importantes sur la fiabilité d'un test et, par extension, sur la manière de les utiliser de manière appropriée.

La première facette concerne les utilisateurs participant au test et leur impact sur la fiabilité de la mesure. Pour les évaluations formatives, de nombreuses études existent et cherchent à trouver le nombre « magique » d'utilisateurs à inclure dans le test pour mettre à jour un nombre suffisant de problèmes critiques (Borsci et al., 2013; Hwang & Salvendy, 2007; Lewis, 2006). Pour les évaluations sommatives, la règle d'usage, basée sur une convention d'usage en statistique appliquée, est d'avoir un échantillon d'au moins de 30 participants. Néanmoins, la probabilité que 30 soit exactement le bon échantillon pour un ensemble de circonstances hétéroclites n'est pas crédible et sa détermination doit plutôt reposer sur la prise en compte de facteurs sources (Lewis, 2014). Statistiquement, on constate que le nombre d'unités statistiques nécessaires pour qu'un test soit positif est inversement proportionnel à sa variance. En d'autres mots, plus les utilisateurs jugent l'utilisabilité d'une interface de manière constante, moins on en a besoin pour déterminer avec fiabilité l'utilisabilité d'une interface. Cette consistance de jugement entre utilisateur pour une situation donnée est appelée fiabilité inter-juges. Le coefficient Kappa et le coefficient de corrélation intra classe (CCI ; Shrout & Fleiss, 1979) sont des mesures communes de la fiabilité inter-juges. Ces outils ont déjà été utilisés dans le domaine de l'évaluation formative, notamment pour contrôler la fiabilité inter-juges d'experts évaluant la gravité d'un problème d'utilisabilité (cf. Jeff Sauro, 2014). De même, des travaux dans le domaine des jeux vidéo ont montré une fiabilité inter-juges faible lors d'évaluations à partir de listes d'heuristiques (Korhonen, Paavilainen, & Saarenpää, 2009; White, Mirza-babaei, McAllister, & Good, 2011) ; et nous mêmes avons montré (Annexe 1) que l'utilisation de critères d'évaluation centrés sur l'expérience utilisateur (plus subjectif) entraîne plus de disparité de

jugement entre les évaluateurs que des critères centrés sur les éléments du produit (plus objectif). De ce fait, cet « *effet évaluateur* » (Hertzum & Jacobsen, 2003; Morten Hertzum, Molich, & Jacobsen, 2014), nous confirme l'importance de disposer de données sur la facette « *utilisateur* », nous permettant de voir la disparité qui peut exister dans une situation donnée, et qui permet ainsi de savoir de quelle ressource humaine il faut disposer pour une évaluation satisfaisante. Néanmoins, ce mode d'évaluation psychométrique bi-facette d'interface (Voir Figure 86) possède l'inconvénient majeur de pouvoir estimer la fiabilité inter-juge que pour un item de mesure à la fois, et donc de ne pas estimer la variabilité d'un item à l'autre, voir les effets d'interactions particuliers que peut présenter un item spécifique.

La deuxième facette importante à prendre en compte concerne les interfaces servant de support aux tests et influant donc sur les résultats des études de validation. Comme la figure 85 nous le montre, la majeure partie des études psychométriques dans le domaine des IHMs se base sur un modèle en deux facettes, en manipulant un échantillon donné d'utilisateurs et d'items. La facette des interfaces, cachée, est fixée artificiellement à un, servant ainsi de référentiel unique pour le test de validation en question. Le risque, c'est de valider/invalider un questionnaire à cause d'un manque de représentativité des interfaces utilisées pour le test. Le même questionnaire pourra, en fonction des écarts plus ou moins grands d'utilisabilité entre les interfaces proposées, constater une sensibilité plus ou moins grande pour un dispositif de mesure exactement similaire. De même, des caractéristiques atypiques de l'interface testée peuvent orienter faussement des mesures de la fiabilité ou de la validité d'un dispositif de mesure. Ce problème de représentativité se retrouve aussi au travers d'une extension récente des questionnaires d'utilisabilité, comme le WAMMI (Kirakowski & Cierlik, 1998), le PSSUQ (Lewis, 2002), le SUPR-Q (Sauro, 2011b) ou le SUS (Bangor et al., 2008; Kortum & Bangor, 2013; Sauro & Lewis, 2012), sous la forme de données normatives. Par exemple, à partir des données de 446 études et de plus de 5000 réponses SUS individuelles, Sauro et Lewis (2012) ont constaté que le score moyen global du SUS est de 68 avec un écart-type de 12,5. Ces données permettent l'interprétation d'un score, en le situant dans la distribution des résultats disponibles. Néanmoins, la limite d'une collecte de données si partielles et disparates est qu'il est impossible de connaitre sa représentativité et de quantifier les facteurs qui jouent sur la variabilité des scores. Par exemple, Sauro et Lewis (2012) constatèrent une différence de moyenne de 8 points pour la distribution totale des scores SUS récoltés dans le cadre de l'étude de Bangor et al. (2008) et de celle de l'étude de Lewis & Sauro (2009). Il faut ainsi garder à l'esprit que ce genre de données est grandement fragilisé par la variation des produit et des tâches incluse dans la base de données (Cavallin, Martin, & Heylighen, 2007), voire également de l'expertise des utilisateurs sollicités. Ainsi, les meilleures références se basent sur des données d'évaluation avec des produits, des tâches, et des utilisateurs similaires (Sauro & Lewis, 2012). En suivant ce principe, Sauro (2011a) organisa les données du SUS en fonction de différents types d'interfaces. Une autre tentative permettant de cerner en profondeur l'impact de la facette « *interface* » dans l'évaluation de l'utilisabilité est portée par Tezza, Bornia & de Andrade (2011). En constatant le déficit des études psychométriques solides, ces derniers utilisèrent la Théorie de Réponse aux Items (TRI), dans l'objectif d'améliorer les méthodologies d'évaluations comparatives de l'utilisabilité. Ils basèrent leur étude sur 361 sites de commerce brésilien choisis aléatoirement à partir d'un moteur de recherche et d'une liste

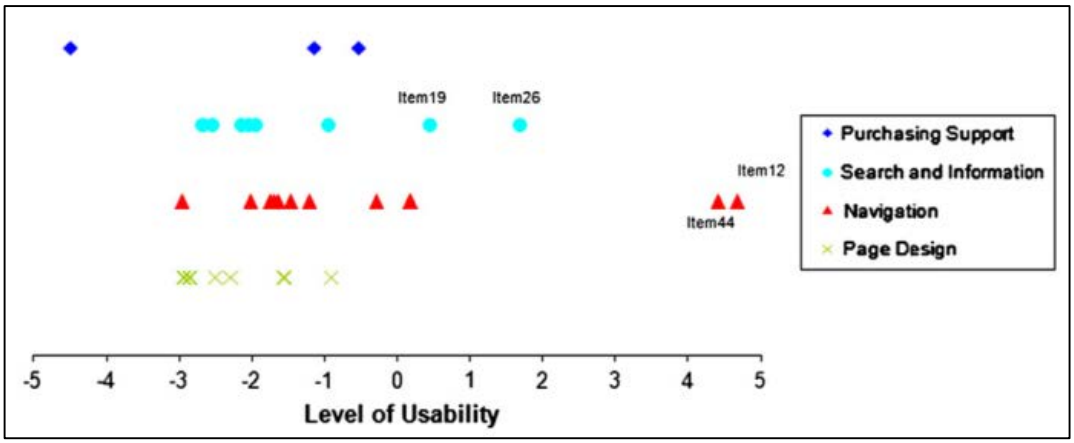


Figure 87 – Items classé en fonction du niveau d'utilisabilité *(tirée de Tezza et al., 2011)*

épurée de 44 items issus de la littérature. L'évaluation a été basée par l'inspection experte d'un des auteurs, sur la totalité des 361 sites et à partir des 44 items sélectionnés. Le résultat d'une telle étude est une analyse riche au niveau de l'item individuel et de leurs combinaisons optimales pour un test donné. Il montre, par exemple, que l'utilisation de l'ensemble de ces items est très efficace pour discriminer des sites internet avec une utilisabilité mauvaise à médiocre (se situant entre -3 et 0), alors que les items 44 et 12, ayant attrait à la navigation, sont les meilleurs du test pour détecter des sites possédant un très haut niveau d'utilisabilité (Figure 87). Néanmoins, ce mode d'évaluation psychométrique bi-facettes (Voir Figure 88) possède également un inconvénient important, celui de faire reposer l'évaluation entière du test sur l'analyse d'un juge unique, et donc de passer à côté l'estimation de la variabilité du jugement, voire de certains effets d'interactions existants entre un évaluateur et certains items de mesures / interfaces.

En conclusion, nous constatons que chacune de ces approches possède sa zone d'ombre, ne permettant pas d'appréhender le modèle minimum de l'évaluation de l'utilisabilité dans toute sa globalité. C'est pourquoi, de plus en plus d'auteurs sont allés au-delà de la théorie classique de la mesure et de la théorie de réponse aux items, en utilisant des modèles multi-facettes. En étendant ainsi le modèle psychométrique de base, l'inclusion de variables (appelées facettes) en supplément de ceux typiquement incluses (les personnes et les items), permet l'amélioration de la complétude et la validité de la mesure. Cette approche est portée, entre autres, par la théorie de la généralisabilité (Cronbach et al., 1972; Cronbach et al., 1963) et les modèles de Rasch multi-facettes (Linacre, 1989).

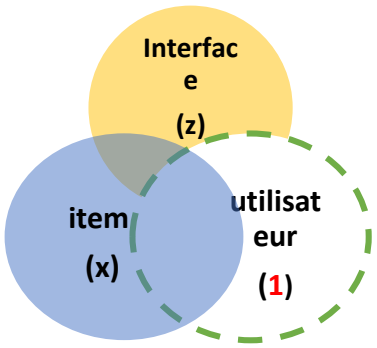


Figure 88 – Illustration en cercles pleins des facteurs pris en compte dans l'étude de Tezza et al. (2011)

Méthodologie

L'objectif de cette étude est de tester la pertinence des modèles d'évaluation multimodale. Par multimodale, nous entendons l'utilisation de mesures combinant des données de sources et de

natures différentes, tel l'usage composé de mesures objectives (ex : taux d'erreur) et subjectives (ex : questionnaire d'utilisabilité perçu). Nous pensons que l'utilisation de données de modalités différentes, par complémentarité, permet de réduire l'impact des biais respectifs de chacune de ces mesures prise en isolation. Pour ce faire, nous avons utilisé un protocole de validation multifacettes, que nous pensons plus adapté que les modèles de validation bi-facettes utilisés communément dans le domaine des IHMs. Nous avons préféré l'utilisation de la théorie de la généralisabilité, à contrario des modèles Rach multi-facettes, car ces derniers présentent comme inconvénient de demander un très grand nombre de participants et d'utiliser de préférence des mesures de nature nominale binaire « Echec/réussite » (Embretson & Reise, 2000). Le cas d'usage choisi pour cette étude est l'évaluation de l'utilisabilité de sites universitaires, à partir d'un outil de capture à distance[88]. Cette partie présente la procédure d'évaluation, ainsi que le plan d'étude utilisé pour les analyses de généralisabilité.

Procédure

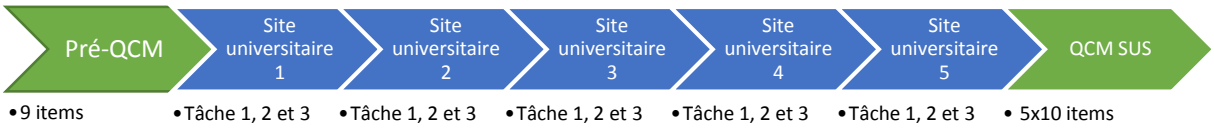


Figure 89 – Ordre de réalisation du test

Plusieurs classes d'étudiants de l'université de Lorraine ont été sollicitées par e-mail (Annexe 1A) pour participer à un test utilisateur à distance sur des sites web universitaires. Le test a consisté, après avoir renseigné un pré-questionnaire (Annexe 1B), à réaliser trois tâches de recherche d'informations, puis à remplir un questionnaire SUS pour chacun des cinq sites universitaires étudiés (Figure 89). Chaque participant a ainsi réalisé 15 tâches et répondu à un questionnaire de 50 items (Annexe 1C). L'ordre d'apparition des cinq sites universitaires et des trois tâches associés a été randomisé pour chacun des participants.

Les 3 tâches ont été sélectionnées pour permettre des activités de difficulté différentes, sur des thèmes assez divers, et pouvant s'effectuer sur n'importe quel site d'université :

- T1 « *Trouver le programme détaillé (avec le volume horaire) d'une formation de M1 donnée* »
- T2 « *Trouver les créneaux horaires d'un sport en particulier dispensé lors des activités extrascolaires* »
- T3 : « *Trouver le lieu et les horaires d'ouverture d'une bibliothèque universitaire en particulier* »

Ces trois tâches ont été personnalisées pour chacun des cinq sites universitaires afin de permettre leur réalisation dans le cadre d'un contenu existant sur chacun de ces sites (Annexe 1D).

[88] Le logiciel de test utilisateur en ligne Evalyzer (www.evalyzer.com)

Figures 90 – Pages d'accueil des sites internet de l'université d'Aix Marseille, de Bretagne Occidentale et de Toulouse 1 Capitole, respectivement

Figures 91 – Pages d'accueil des sites internet de l'université du Sud Toulon-Var et de Caen, respectivement

Les cinq sites universitaires ont été sélectionnés par tirage au sort dans la liste des 82 sites internet des universités françaises existantes, afin d'avoir un échantillon dont la variabilité en termes de niveau d'utilisabilité soit la plus représentative possible. Les sites internet qui ont été sélectionnés sont associés aux universités d'Aix-Marseille (http://www.univ-amu.fr/), de Caen (http://www.unicaen.fr/), du Sud Toulon-Var (http://www.univ-tln.fr/), de Bretagne Occidentale (http://www.univ-brest.fr/) et de Toulouse 1 Capitole (http://www.univ-tlse1.fr/); (Figures 90 et 91).

Mesures

L'étude a recueilli de nombreux indicateurs de l'utilisabilité, utilisés sous leurs formes brutes, standardisés ou combinés, avec l'objectif d'analyser pour chacun leurs caractéristiques psychométriques associées. Il a été décidé de choisir des indicateurs selon deux critères. Le premier critère est la facilité de recueil des mesures. En effet, un des reproches fait à l'approche multimodale est sa lourdeur en termes de collecte de données, qui réduit le bénéfice associé à l'augmentation de la qualité de la mesure. Nous avons donc choisi des mesures faciles et rapides à collecter, voire pouvant être récupérées de manière automatisée. Le deuxième critère est la diversité des mesures sélectionnées. Nous visons ainsi à prouver l'avantage d'utiliser des mesures issues de modalités différentes afin de disposer, par leurs complémentarités, d'une procédure d'évaluation plus solide.

Les données issues du questionnaire SUS ont permis premièrement de disposer d'un **indicateur subjectif** de l'utilisabilité pour chacun des sites universitaires, au travers de son évaluation par chacun des participants à l'étude. Le SUS a été utilisé pour ces avantages en termes de vitesse d'administration et de ses qualités psychométriques maintes fois éprouvées. Les 10 items du questionnaire ont été combinés, puis standardisés sous un score compris entre 0 et 1.

De plus, de nombreuses données comportementales ont été capturées, telles que le temps de réalisation des tâches, la distance à la souris, le nombre de clics, le nombre d'URL visités, le taux de revisite et la réussite des tâches. Toutes ces données ont été capturées automatiquement par l'outil en ligne *Evalyzer*, pour tous les sites universitaires testés et toutes les tâches réalisées. Ces données ont été ensuite standardisées et combinées pour mettre au point différents indicateurs composites de l'utilisabilité. Un **indicateur de performance** a été mis au point en utilisant le temps de réalisation des tâches, la distance à la souris, le nombre de clics, le nombre d'URL visités et le taux de revisite. Afin de pouvoir les combiner, ces données ont été normalisées[89] individuellement pour obtenir des scores variant entre 0 et 1. Pour ne pas réduire la sensibilité des mesures par l'impact des valeurs extrêmes sur la normalisation, des scores limites ont été déterminés. La spécification de ces scores limites peut être déterminée en utilisant un score de référence, une échelle absolue (Sauro & Kindlund, 2005) ou encore un multiple du score modale (Kainda, Flechais, & Roscoe, 2009). Pour cette étude, les dix valeurs de l'extrémité base ont été plafonnées par la plus grande de ces valeurs et les dix valeurs de l'extrémité haute ont été plafonnées par la plus petite de ces valeurs. Seul le taux de revisite n'a pas été normalisé, car variant naturellement entre 0 et 1. Ces données ont ensuite été inversées, pour varier dans le sens d'une utilisabilité grandissante, puis ont été fusionnées en un seul indicateur de performance**.**

L'**indicateur de réussite**, représenté par une valeur brute de 0 ou 1, correspond à la réussite de la tâche par l'utilisateur et pour tous les sites universitaires testés. La réussite à la tâche est détectée automatiquement par Evalyzer en comparant la page d'arrivée avec la ou les pages préenregistrées dans l'application comme une réussite du scénario.

Deux autres indicateurs ont également été construits :

- L'**indicateur objectif**, composé de la moyenne de l'indicateur de réussite et de performance, et calculé pour chacune des tâches et des sites universitaires ;
- **L'indicateur composite général**, composé de la moyenne de l'indicateur subjectif, de réussite et de performance (cf. Tullis & Albert, 2008), et calculé pour chacun des sites universitaires

Plans d'étude pour l'analyse de la généralisabilité

Un plan d'étude a été mis en place afin d'estimer la fiabilité du protocole du test selon la théorie de la généralisabilité. Il a consisté à identifier les facettes d'intérêts, la nature de leurs relations et leurs types d'échantillonnages. Puis, un objet d'étude a été déterminé, ainsi que le type de mesure visée par le protocole de mesure développé. Ce plan a ensuite été décliné entre deux études G et plusieurs études D pour prendre en compte l'asymétrie des données (les données

[89] Selon la formule : $\frac{x-min}{max-min}$

objectives ont été récoltées pour les tâches alors que les données subjectives n'ont été récoltées qu'au niveau des sites universitaires.

Dans notre étude, les facettes « *Site universitaire* », « *Tâche* », « *Utilisateur* » et « *Indicateur* » ont été sélectionnées. La facette « Site universitaire » représente l'ensemble des sites internet universitaires que nous cherchons à évaluer et à discriminer, en mettant au point ce protocole de mesure. La facette « *Tâche* » contient les différentes tâches réalisées par les utilisateurs lors du test utilisateur. C'est un point intéressant à étudier lors de l'évaluation de l'utilisabilité d'une interface (Cavallin et al., 2007), car leur sélection peut influer sur la difficulté d'un test. La facette « *Utilisateur* » contient les participants ayant réalisé les tâches et rempli les questionnaires d'utilisabilité. L'étude de cette facette nous permet de quantifier la part de variabilité comportementale et de subjectivité individuelle contenue dans ce dispositif d'évaluation. Enfin, la facette « *Indicateur* » contient les différents indicateurs utilisés lors du test utilisateur. Dans cette étude, toutes les facettes ont été croisées ensemble, selon un plan d'observation : S x F x R x J (Figure 92). Ainsi, tous les utilisateurs ont réalisé toutes les tâches sur tous les sites universitaires.

Tableau 12 – Univers d'échantillonnage et niveau des facettes pour l'étude G

Abrév.	Nom de la facette	Niveau	Univers
S	Site Universitaire	5	Infini
U	Utilisateur	30	Infini
T	Tâche	3 (Etude G1) 0 (Etude G2)	Infini (Etude G1) Caché (Etude G2)
I	Indicateur	2 (Etude G1) 3 (Etude G2)	2 (Etude G1) 3 (Etude G2)

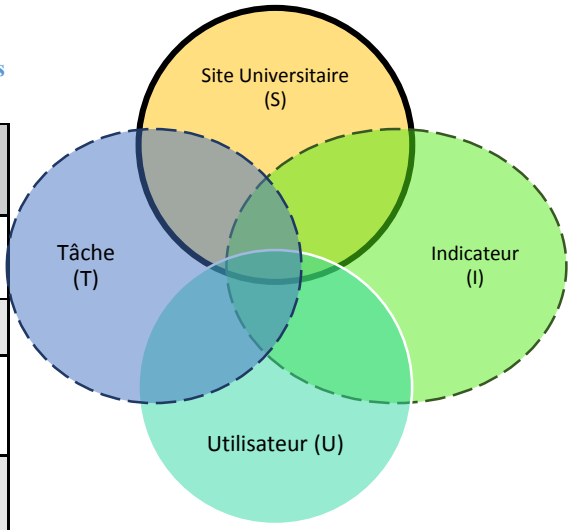


Figure 92 – Plan d'observation SUTI, modulées en fonction des études G

Les univers d'échantillonnage pour les facettes « *Site Universitaire* », « *Utilisateur* » et « *Tâche* » ont été spécifiés comme infinis. Il s'agira donc de tester la fiabilité d'un protocole d'évaluation dont les éléments qui le composent, c'est-à-dire les sites universitaires testés, les tâches évaluées et les participants sollicités, varient d'un test à l'autre. Par contre, l'univers d'échantillonnage de la facette « *Indicateur* » a été fixé à trois. Il s'agira donc de restreindre l'étude des indicateurs de mesure de l'utilisabilité à ceux utilisés dans le test, c'est-à-dire aux indicateurs subjectifs, de réussite, de performance et leurs combinaisons. Vu que l'indicateur subjectif a été recueilli par site universitaire et non pour chacune des tâches, l'analyse de la généralisabilité a été réalisée en deux études G. Le niveau respectif de chacune des facettes correspondant à l'étude G1 et G2 est reporté dans le tableau 12. L'étude G1 contient l'analyse des quatre facettes, en enlevant le niveau « *subjectif* » de la facette « *indicateur* ». Cette étude G et les études D associées, permettent d'analyser la pertinence des indicateurs de réussite, de performance et leurs combinaisons (indicateur objectif), ainsi que l'impact du nombre de tâches et d'utilisateurs sur la fiabilité de la mesure. L'étude G2 contient l'analyse des trois facettes : «

Site Universitaire », « *Utilisateur* » et « *Indicateur* ». La facette « *Tâche* » a été cachée, mais le niveau « *subjectif* » de la facette « *Indicateur* » a été ajouté, permettant ainsi de compléter l'analyse de l'étude G1 par l'analyse de la pertinence des indicateurs subjectifs, ainsi qu'un indicateur composite général.

Comme le protocole d'évaluation vise à comparer divers sites universitaires entre eux, la facette « *Site Universitaire* » a été choisie comme facette de différenciation. Les facettes « *Utilisateur* », « *Tâche* », et « *Indicateur* » ont été désignées comme facettes de mesure. Pour l'étude G, les coefficients de généralisabilité absolus ont été calculés, accompagnés de leurs erreurs standard de mesure. En effet, ce coefficient évalue la capacité d'un test à déterminer la position exacte de chaque entité sur une l'échelle de mesure (ex : un élève obtient une note de 15 sur 20), indépendamment des autres individus. Cette capacité est pratique dans le milieu de l'utilisabilité car elle réduit les coûts, les participants n'ayant pas l'obligation de tester toutes les interfaces lors d'une évaluation de divers produits numériques.

Résultats

Statistiques descriptives et corrélations entre les mesures

L'évaluation a été réalisée à partir de 30 utilisateurs, dont 17 hommes (56,7%) et 13 femmes (43,3%). L'âge moyen était de 23,8 ans ($\sigma = 4,25$), réparti entre élèves de master 2 (20, soit 70%) et de licence 2 (10, soit 30%). 56,7% (n = 17) des utilisateurs déclarent se connecter sur le site de leur université sur une base journalière, 26,7% (n = 8) sur une base hebdomadaire, et 16% (n = 5) moins d'une fois par semaine. Dans 4 cas sur 150, un des élèves s'est connecté sur un des cinq sites testés une fois seulement. De même, dans 3 cas sur 150, un des élèves s'est connecté sur un des cinq sites testés plusieurs fois. Cependant, aucun effet significatif sur l'indicateur composite général d'utilisabilité n'a été observé pour le fait de s'être connecté une ou plusieurs fois sur ces sites ($t(142) = 0,231$, $p = .82$).

Les corrélations entre les différentes mesures de l'utilisabilité nous renseigne sur leur niveau de dépendance en termes d'informations récoltées sur l'utilisabilité des sites universitaires (Tableau 13). On constate que les mesures individuelles de performance ne sont pas ou peu liées à l'indicateur de réussite aux tâches (Tableau 13, valeurs en rouge). En effet, la majorité des corrélations ne sont pas statistiquement significatives ($p > .05$) et seule une corrélation négative entre l'indicateur de réussite aux tâches et de la distance du parcours à la souris est significative au seuil alpha de 5% ($r = -0,172$, $p < .05$). Par contre, la plupart des mesures individuelles de performance sont fortement corrélées entres elles (Tableau 13, valeurs en orange), affichant des coefficients de Pearson compris entre 0,596 et 0,914 ($p > .01$). Cela nous montre que les mesures individuelles de performance et l'indicateur de réussite aux tâches nous donnent des informations différentes sur l'utilisabilité d'un produit. On constate également que la mesure subjective de l'utilisabilité, représentée par le questionnaire SUS, est corrélée moyennement avec mesures individuelles de performance (Tableau 13, valeurs en verte, corrélations comprises entre -0262 et -0,436, p>.01), et l'indicateur de réussite (Tableau 13, valeurs en violète, corrélations $r = 0,456$, $p < .01$).

Tableau 13 – Corrélations entre les mesures de performance, de réussite et subjectives (SUS)

	Réussite	Temps	Clics souris	Distance souris	URL visités	Taux de revisite
Temps	-0,079					
Clics souris	-0,081	0,867**				
Distance souris	-0,172*	0,889**	0,856**			
URL visités	-0,042	0,859**	0,914**	0,850**		
Taux de revisite	-0,101	0,622**	0,670**	0,616**	0,596**	
SUS	0,456**	-0,415**	-0,390**	-0,436**	-0,347**	-0,262**

* p < .05, ** p < .01

Etude G1 : analyse des indicateurs objectifs

Le tableau 14 expose l'analyse de la variance et sa décomposition en divers éléments dans le cadre de l'étude G1. Il montre que la variation engendrée par l'interaction entre le site universitaire, la tâche et l'utilisateur est particulièrement importante (19,6%), tout comme la variabilité résiduelle, non expliquée, (35,5%). L'interaction entre l'utilisateur et le type d'indicateur objectif (performance ou réussite) est également important (14,7%). En comparaison, la variation de différenciation est de 12,5%. Ce qui veut dire que pour que le dispositif de mesure soit assez fiable, il conviendra de jouer sur ces paramètres, c'est à dire sur l'impact des facettes « Indicateur », « Utilisateur » et « Tâche ».

Tableau 14 – ANOVA et calcul des composants de la variance pour l'étude G1

Source de variations	Somme des carrés	ddl	Carré moyen	Composants				
				Aléatoire	Mixte	Corrigé	%	Erreur standard
S	20.81330	4	5.20333	0.02115	0.02655	0.02655	12.5	0.01711
I	2.20848	1	2.20848	0.00051	0.00051	0.00026	0.1	0.00435
T	0.92142	2	0.46071	-0.00115	0.00019	0.00019	0.1	0.00175
U	6.84826	29	0.23615	-0.01021	0.00532	0.00532	2.5	0.00520
SI	4.49995	4	1.12499	0.01079	0.01079	0.01079	5.1	0.00725
ST	3.34347	8	0.41793	0.00476	0.00559	0.00559	2.6	0.00326
SU	10.43979	116	0.09000	-0.00367	0.00121	0.00121	0.6	0.00345
IT	1.02288	2	0.51144	0.00267	0.00267	0.00267	1.3	0.00244
IU	16.16046	29	0.55726	0.03107	0.03107	0.03107	14.7	0.00952
TU	4.01486	58	0.06922	-0.00003	-0.00135	-0.00135	0.0	0.00199
SIT	0.99652	8	0.12456	0.00165	0.00165	0.00165	0.8	0.00187
SIU	12.10956	116	0.10439	0.00976	0.00976	0.00976	4.6	0.00509
STU	19.19352	232	0.08273	0.00381	0.04137	0.04137	19.6	0.00517
ITU	3.59305	58	0.06195	-0.00263	-0.00263	-0.00263	0.0	0.00265
SITU	17.42524	232	0.07511	0.07511	0.07511	0.07511	35.5	0.00694
Total	123.59076	899					100%	

En se basant sur les données du tableau précédent, le tableau 15 présente les résultats de l'étude G1 pour le plan S/ITU. Le coefficient de généralisabilité absolu obtenu lors de cette étude est élevé (Φ = .91), ce qui traduit une bonne fiabilité du dispositif à déterminer la position exacte de chaque site universitaire sur l'échelle objective de mesure de l'utilisabilité (de 0 à 1). Ce bon score s'explique en premier par le nombre satisfaisant de tâches (3) et d'utilisateurs (30) qui représente 89,3% de la variance d'erreur absolue. Le deuxième facteur qui explique ce bon score est d'avoir fait le choix de fixer la facette indicateur, ce qui permet de retirer sa variance d'erreur au protocole de mesure. Néanmoins cela implique que les praticiens voulant utiliser ce

protocole de mesure, doivent impérativement utiliser dans chacun de leur test les mêmes éléments inclus dans la facette indicateur, c'est-à-dire ici l'indicateur de performance et de réussite.

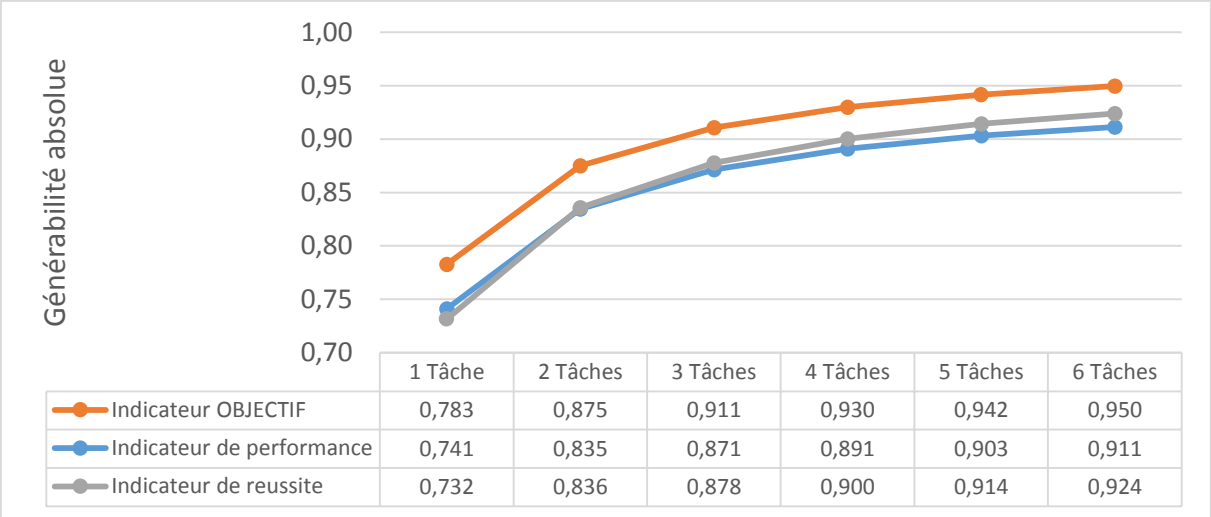
La facette Utilisateur, contribue à 6,8% de la variance de mesure, ce qui est plutôt faible. Cela correspond à la variabilité de performance/réussite d'un participant à l'autre. Si l'on regarde les facettes d'interactions SU ou TU, on constate également que le niveau de performance d'un utilisateur ne varie pas beaucoup, si l'on considère seulement les tâches ou les sites universitaires (1,6% de la variance totale). Par contre, les performances d'un participant peuvent varier fortement en fonction d'une/des tâche(s) effectuées sur un site en particulier, car la facette d'interaction STU représente 17,7% de la variance totale.

Enfin, même si les tâches en elles-mêmes ne diffèrent pas beaucoup en termes de difficulté (elles ne génèrent que 2,4% de variance d'erreur), cette variabilité augmente fortement en fonction des sites universitaires. En effet, 71,6% de la variance d'erreur absolue provient de l'interaction entre la facette site universitaire et tâche, ce qui confirme l'intérêt de tester l'utilisabilité d'un produit sur plus d'une tâche.

Tableau 15 – Répartition de la variance et calcul des coefficients de généralisabilité pour l'étude G1

Variance de différenciation		**Variance d'erreur**				
Source	**Variance**	**Source**	**Variance d'erreur relative**	**%**	**Variance d'erreur Absolue**	**%**
S	0.02655		.....		.....	
	.....	I	.....		(0.00000)	0.0
	.....	T	.....		0.00006	2.4
	.....	U	.....		0.00018	6.8
	.....	SI	(0.00000)	0.0	(0.00000)	0.0
	.....	ST	0.00186	78.8	0.00186	71.6
	.....	SU	0.00004	1.7	0.00004	1.6
	.....	IT	.....		(0.00000)	0.0
	.....	IU	.....		(0.00000)	0.0
	.....	TU	.....		(0.00000)	0.0
	.....	SIT	(0.00000)	0.0	(0.00000)	0.0
	.....	SIU	(0.00000)	0.0	(0.00000)	0.0
	.....	STU	0.00046	19.5	0.00046	17.7
	.....	ITU	.....		(0.00000)	0.0
	.....	SITU	(0.00000)	0.0	(0.00000)	0.0
Somme de la variance	0.02655		0.00236	100%	0.00260	100%
Écart-type	0.16293		**Écart-type relatif :**	0.04860	**Écart-type absolu :**	0.05101
Ep^2	0.92					
Φ	0.91					

En manipulant la facette « Indicateur », les études D menées nous permettent de comparer le coefficient de généralisabilité absolu en fonction des indicateurs de mesure de l'utilisabilité retenus dans le protocole de test. On constate ainsi que l'indicateur de réussite et de performance présente un coefficient de généralisabilité très proche. En effet, par exemple, pour un protocole de test composé de 3 tâches et de 30 utilisateurs, Φ = 0,871 pour l'utilisation de l'indicateur de performance seul, et Φ = 0,878 pour l'utilisation de l'indicateur de réussite seul, soit une différence de moins d'1%. Néanmoins, dans le cas d'un protocole avec peu d'utilisateurs (ex : 5 ou 10), on constate que l'utilisation d'un indicateur de réussite présente un avantage sur

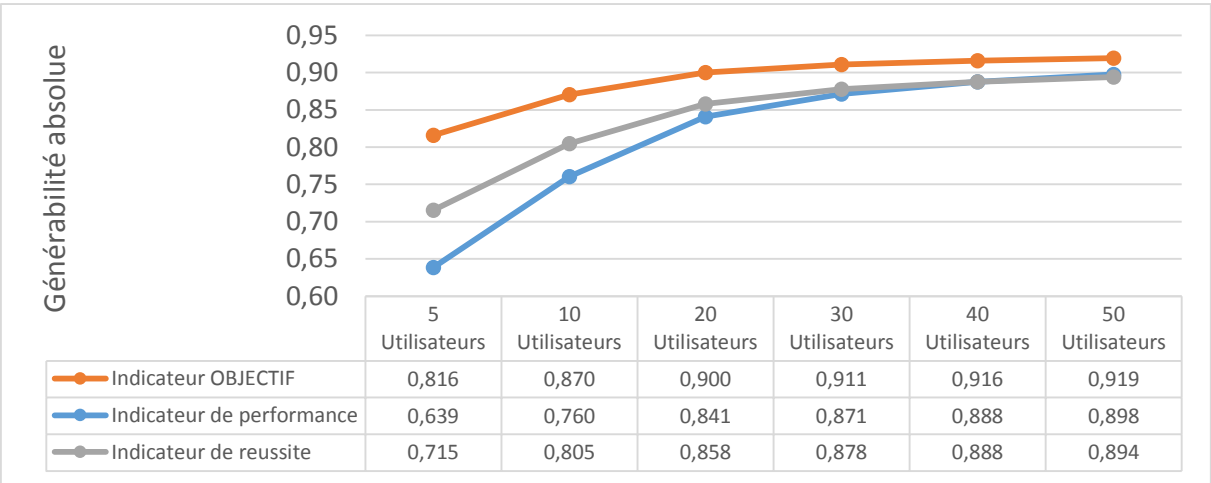


	1 Tâche	2 Tâches	3 Tâches	4 Tâches	5 Tâches	6 Tâches
Indicateur OBJECTIF	0,783	0,875	0,911	0,930	0,942	0,950
Indicateur de performance	0,741	0,835	0,871	0,891	0,903	0,911
Indicateur de reussite	0,732	0,836	0,878	0,900	0,914	0,924

Figure 93 – Etude D sur l'évolution du coefficient de généralisabilité absolue en fonction du nombre de tâches et du type d'indicateur de mesure (la facette « Utilisateur » est composée de 30 éléments)

l'indicateur de performance, cet écart s'effaçant rapidement à partir de 20 utilisateurs (Figure 94). On perçoit également une meilleur fiabilité du protocole de mesure quand on utilise ces indicateurs ensemble, qu'isolement. En effet, les coefficients de généralisabilité absolus pour l'indicateur objectif surpassent toutes les valeurs obtenues pour les indicateurs de performance et de réussite. Cet écart est important quand le nombre d'éléments des facettes « Utilisateur » et « tâche » est faible et ne plafonne pas quand le nombre d'éléments pour ces derniers est fort (figure 93 et 94). Cela montre que l'avantage d'utiliser un indicateur composite de l'utilisabilité, comprenant la réussite et la performance à une tâche, ne peut pas être compensé par l'utilisation d'un grand nombre d'utilisateurs ou de tâches.

De plus, les études D menées sur la variation du coefficient de généralisabilité absolu confirment les observations sur la contribution de la facette « Tâche » et « Utilisateur » sur la fiabilité de la mesure (Figure 93 et 94). On constate ainsi un effet important du nombre de tâches sur le coefficient de généralisabilité absolu : pour un protocole de test à 30 utilisateurs utilisant l'indicateur objectif d'utilisabilité, $\Phi = 0.783$ pour 1 tâche, 0,911 pour 2 tâches, et 0,950 pour 6 tâches. Le nombre d'utilisateurs participant au test a également un impact non négligeable sur la qualité de la mesure : pour un protocole de test à 3 tâches utilisant l'indicateur



	5 Utilisateurs	10 Utilisateurs	20 Utilisateurs	30 Utilisateurs	40 Utilisateurs	50 Utilisateurs
Indicateur OBJECTIF	0,816	0,870	0,900	0,911	0,916	0,919
Indicateur de performance	0,639	0,760	0,841	0,871	0,888	0,898
Indicateur de reussite	0,715	0,805	0,858	0,878	0,888	0,894

Figure 94 – Etude D sur l'évolution du coefficient de généralisabilité absolue en fonction du nombre d'utilisateur et du type d'indicateur de mesure (la facette « Tâche » est fixée à 3 éléments)

objectif d'utilisabilité, Φ = 0.870 pour 10 utilisateurs, 0,900 pour 20 utilisateurs, et 0,916 pour 40 utilisateurs. On constate ainsi un écart de fiabilité important quand très peu d'utilisateurs participent au test, mais cet écart s'efface rapidement après 20 utilisateurs. Pour augmenter la fiabilité du test, il est donc préférable d'ajouter des tâches plutôt que des utilisateurs, quand ces derniers dépassent les 20.

Etude G2 : étude des rapports entre indicateurs objectifs et subjectifs

Les tableaux 16 et 17 exposent la décomposition de la variance et le calcul du coefficient de généralisabilité dans le cadre de l'étude G2. Cette étude incorpore l'indicateur subjectif dans le calcul général d'une mesure de l'utilisabilité composite. Cette opération nous oblige toutefois à cacher la facette «tâche » de notre analyse et à utiliser un plan simplifié S/IU, car nous n'avons pas de donnée subjective au niveau de la tâche unitaire.

Tableau 16 – ANOVA et calcul des composants de la variance pour l'étude G2

Source de variations	Somme des carrés	ddl	Carré moyen	Composants				
				Aléatoire	Mixte	Corrigé	%	Erreur standard
S	11.59025	4	2.89756	0.02849	0.03179	0.03179	28.7	0.01866
I	4.25451	2	2.12726	0.01146	0.01146	0.00764	6.9	0.01008
U	2.27554	29	0.07847	-0.00268	0.00277	0.00277	2.5	0.00194
SI	2.61539	8	0.32692	0.00990	0.00990	0.00990	8.9	0.00487
SU	4.27941	116	0.03689	0.00231	0.01230	0.01230	11.1	0.00185
IU	6.47896	58	0.11171	0.01635	0.01635	0.01635	14.8	0.00412
SIU	6.95266	232	0.02997	0.02997	0.02997	0.02997	27.1	0.00277
Total	38.44672	449					100%	

Tableau 17 – Répartition de la variance et calcul des coefficients de généralisabilité pour l'étude G2

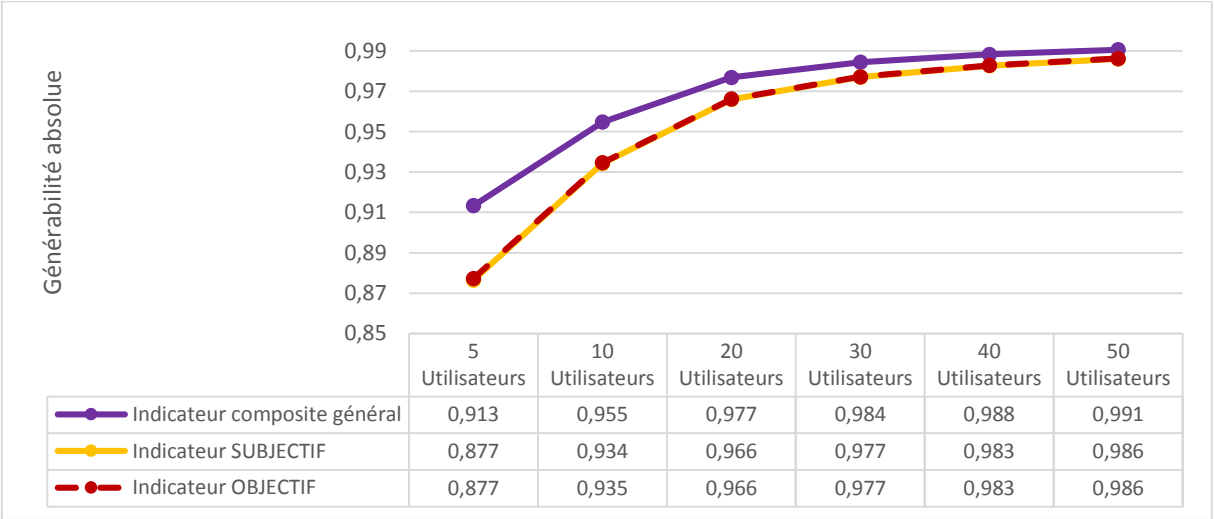
Variance de différenciation		Variance d'erreur				
Source	Variance	Source	Variance d'erreur relative	%	Variance d'erreur Absolue	%
S	0.03179		.....		.....	
	.....	I	.....		(0.00000)	0.0
	.....	U	.....		0.00009	18.4
	.....	SI	(0.00000)	0.0	(0.00000)	0.0
	.....	SU	0.00041	100.0	0.00041	81.6
	.....	IU	.....		(0.00000)	0.0
	.....	SIU	(0.00000)	0.0	(0.00000)	0.0
Somme de la variance	0.03179		0.00041	100%	0.00050	100%
Écart-type	0.17828		**Écart-type relatif :** 0.02025		**Écart-type absolu :**	0.02241
Ep^2	0.99					
Φ	0.98					

Le coefficient de généralisabilité absolu obtenu lors de cette étude est très élevé (Φ = .98). Même si l'ajout de l'indicateur subjectif peut améliorer en partie la fiabilité de la mesure, on constate ici l'impact du masquage de la facette « Tâche » sur l'estimation du coefficient généralisabilité. En effet, nous avons vu que cette facette est le premier facteur produisant directement ou indirectement la plus grande partie de la variance d'erreur du protocole de mesure. Son retrait implique que le modèle ne contient plus que les facettes « Utilisateur » et

« Indicateur », pouvant générer de l'erreur de mesure. Il conviendra donc de ne pas s'attacher aux valeurs absolues des coefficients de généralisabilité mais aux rapports relatifs existant entre les différents types d'indicateurs de l'utilisabilité testés.

En manipulant la facette « Indicateur », les études D menées nous permettent de comparer les coefficients de généralisabilité absolus en fonction des indicateurs de mesure de l'utilisabilité retenus dans le protocole de test. On constate ainsi que l'indicateur subjectif et objectif présente un coefficient de généralisabilité pratiquement similaire. En effet, par exemple, pour un protocole de test composé de 3 tâches et de 30 utilisateurs, $\Phi = 0{,}871$ pour l'utilisation de l'indicateur de performance seul, et $\Phi = 0{,}878$ pour l'utilisation du l'indicateur de réussite seul, soit une différence de moins d'1%.

On perçoit également une meilleur fiabilité du protocole de mesure quand on utilise tous ces indicateurs ensemble, qu'isolément (Figure 95). En effet, les coefficients de généralisabilité absolus pour l'indicateur composite général surpassent toutes les valeurs obtenues pour les indicateurs objectif et subjectif. Quel que soit le nombre d'utilisateur participant à l'étude, cette combinaison enlève un tiers de l'écart séparant les mesures individuelles d'une mesure parfaitement fiable (c'est-à-dire $\Phi = 1$).

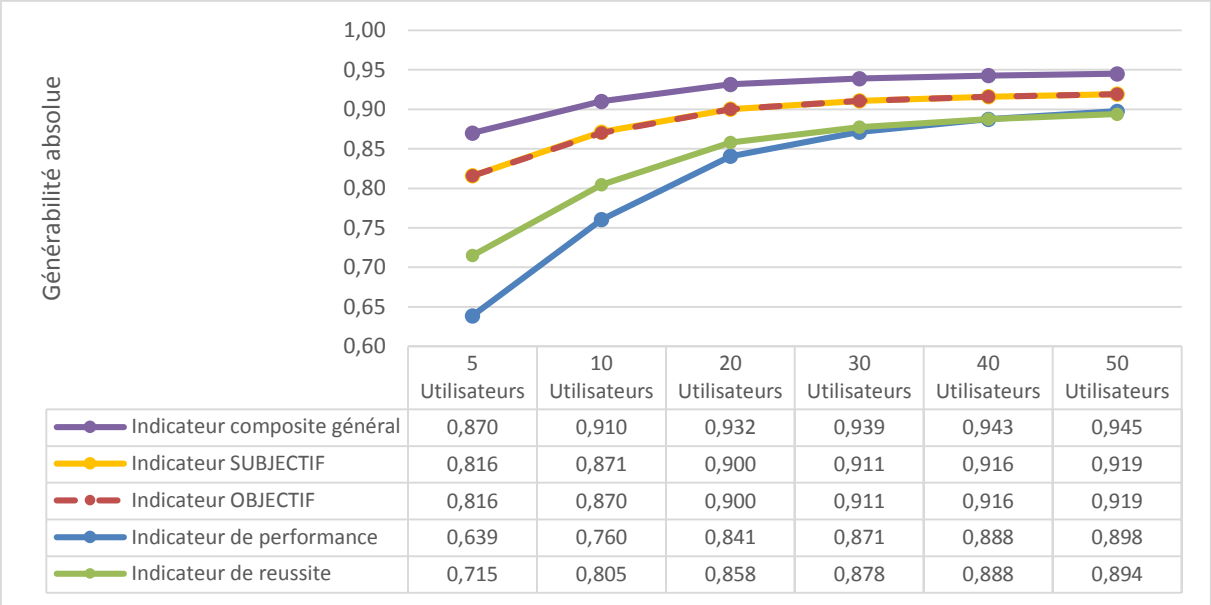


	5 Utilisateurs	10 Utilisateurs	20 Utilisateurs	30 Utilisateurs	40 Utilisateurs	50 Utilisateurs
Indicateur composite général	0,913	0,955	0,977	0,984	0,988	0,991
Indicateur SUBJECTIF	0,877	0,934	0,966	0,977	0,983	0,986
Indicateur OBJECTIF	0,877	0,935	0,966	0,977	0,983	0,986

Figure 95 – Etude D sur l'évolution du coefficient de généralisabilité absolu en fonction du nombre d'utilisateur et du type d'indicateur de mesure (la facette « Tâche » est cachée à 3 éléments)

Conclusion

Cette étude nous renseigne sur de nombreux éléments nous permettant d'améliorer nos pratiques en termes de mesure de l'utilisabilité dans le contexte des tests utilisateurs sommatifs. Le premier apport concerne la fiabilité relative de nombreuses mesures et composés couramment utilisés dans les évaluations de l'utilisabilité. En effet, cette étude a permis de démontrer l'intérêt des indicateurs composites, agrégeant des informations de différentes modalités et basées sur des données récupérées automatiquement par le système, pour l'évaluation de l'utilisabilité.



	5 Utilisateurs	10 Utilisateurs	20 Utilisateurs	30 Utilisateurs	40 Utilisateurs	50 Utilisateurs
Indicateur composite général	0,870	0,910	0,932	0,939	0,943	0,945
Indicateur SUBJECTIF	0,816	0,871	0,900	0,911	0,916	0,919
Indicateur OBJECTIF	0,816	0,870	0,900	0,911	0,916	0,919
Indicateur de performance	0,639	0,760	0,841	0,871	0,888	0,898
Indicateur de reussite	0,715	0,805	0,858	0,878	0,888	0,894

Figure 96 – Synthèse des Etudes D sur l'évolution du coefficient de généralisabilité absolu en fonction du nombre d'utilisateur et du type d'indicateur de mesure (la facette « Tâche » est cachée à 3 éléments)

La Figure 96, présente une synthèse de ces résultats en corrigeant les estimations de la généralisabilité absolue pour les indicateurs subjectifs et composite général[90]. Il démontre, premièrement, l'intérêt de combiner divers composants objectifs de la mesure de l'utilisabilité, telle que la réussite à une tâche (composant d'efficacité) et la performance à une tâche (composant d'efficience). Cette fusion augmente la finesse de la mesure, en gommant le biais propre à la différence de persévérance des utilisateurs. En effet, compter seulement sur la mesure de la réussite à la tâche, c'est être incapable de prendre en compte les abandons rapides ou encore les réussites coûteuse en énergie. Compter seulement sur la mesure de la performance à la tâche (c'est-à-dire sur les ressources dépensées), c'est ne pas tenir compte du résultat de la tâche réalisée, positif ou négatif. Au contraire, en combinant ces deux mesures, il est possible d'analyser plus finement les résultats, tel qu'illustré schématiquement avec l'échelle suivante :

1. **Score de 1 :** réussite rapide (indicateur de réussite : 1 ; indicateur de performance : 1)
2. **Score de 0,5 :** abandon rapide (indicateur de réussite : 0 ; indicateur de performance : 1) ou réussite laborieuse (indicateur de réussite : 0 ; indicateur de performance : 1)
3. **Score de 0 :** échec laborieux (indicateur de réussite : 0 ; indicateur de performance : 1)

Ainsi, l'utilisation de cet indicateur composite étend les indicateurs individuels au-delà de leurs biais respectifs, de leurs zones d'ombres faisant gagner l'analyse en contraste et sensibilité. Le même résultat est à l'œuvre avec la fusion des indicateurs objectifs et subjectifs. D'une fiabilité constatée quasi-identique entre ces deux indicateurs séparés, leur fusion nous offre un composé d'une précision significativement plus grande, quelles que soient les conditions des protocoles de mesures simulées lors des études D. Cela nous montre que ces deux modalités d'évaluation sont complémentaires et que leurs utilisations conjointes est bénéfique pour la mesure de l'utilisabilité. De manière plus générale, nous constatons que la combinaison d'indicateurs pointant sur un construit commun, peu corrélés en eux, mais de fiabilité proche, augmente la

[90] La correction appliquée a été calculée en introduisant l'impact de la facette « Tâche » sur les données et en gardant les rapports géométriques entre l'écart des indicateurs à corriger avec les scores des indicateurs objectifs, d'une part, et l'écart avec le score parfait (1), d'autre part.

fiabilité de l'indicateur ainsi combiné. Ainsi, l'indépendance des sources nous permet de neutraliser certains biais inhérents à l'acquisition de certaines données et d'obtenir une mesure globale plus fiable. Cela va à contre sens de la théorie classique de la mesure (en tout cas dans sa traduction la plus courante et étroite), qui considère que la fiabilité de la mesure est d'autant plus importante que ces items sont proches les uns des autres. Au contraire, cela nous rapproche de la théorie de réponse aux items, qui voit, dans chacun des éléments d'un test une quantité d'informations nouvelles plus ou moins importantes, portées par chacun d'entre eux, et nourrissant la sensibilité générale du test global.

Le deuxième apport important de cette étude concerne la pertinence des protocoles multifacettes dans le domaine particulier des interactions homme-machine et dans le contexte de l'évaluation à distance. En effet, cette étude a permis de démontrer l'intérêt de cette approche statistique, qui permet d'estimer l'influence de certains paramètres agissant sur la fiabilité de la mesure dans le contexte de l'évaluation du numérique et couramment occulté dans les recherches méthodologiques actuelles. Nous avons ainsi constaté l'impact de certains facteurs, comme le nombre de tâches et d'utilisateurs ayant servi pour le test. Nous avons également vu qu'il est possible d'estimer précisément le nombre d'éléments nécessaires dans chaque facette pour que le test soit considéré comme suffisamment fiable, en se basant sur leurs instabilités respectives. Ainsi, nous avons observé que les performances et l'utilisabilité perçues varient en quantités importantes en fonction de la tâche, bien plus encore que d'un utilisateur à l'autre. Il convient donc de ne pas négliger cette facette lors de la mise en place d'un protocole de test. La facette utilisateur, souvent considérée comme importante pour la fiabilité d'un test à également pu être examinée précisément. En effet, l'avantage de la théorie de la généralité est de pouvoir chiffrer cet effet sur la fiabilité de la mesure, allant même jusqu'à déterminer le nombre d'utilisateurs nécessaire au test pour bénéficier d'un coefficient de généralisabilité donné. Ainsi, les résultats de notre étude nous informe qu'il est possible d'estimer de manière convenable ($\Phi = 0{,}858$) l'utilisabilité de plusieurs sites universitaires à partir du taux de réussite à trois tâches différentes par un échantillon d'au moins 20 utilisateurs. De même, le nombre d'utilisateurs nécessaires peut diminuer lorsque l'on utilise une mesure de l'utilisabilité plus complète (en prenant en compte la performance à la tâche ou l'avis subjectif des utilisateurs) ou que l'on ajoute des tâches supplémentaires à réaliser lors du test.

Cependant, une des limites de l'étude est d'avoir séparé l'estimation de la facette tâche et de la mesure subjective de l'utilisabilité. En effet, afin de ne pas demander à l'utilisateur de remplir le questionnaire SUS 15 fois de suite, nous avons limité cette tâche à 5, une pour chaque site universitaire, au moment de la fin du test. C'est une des concessions qui a été décidée afin de contourner une des limites des protocoles de validation utilisant la théorie de la généralisabilité : la longueur des passassions par participant. Utiliser un questionnaire très court, à la fin de chaque tâche, aurait pu être également une alternative.

En effet, l'évaluation multimodale, utilisant des indicateurs complémentaires, permet de contrebalancer la perte individuelle de fiabilité par une meilleure validité des mesures. Ces propriétés particulièrement intéressantes pour interroger des entités complexes, tels que les émotions ou l'engagement, seront utilisées dans l'étude suivante. Elle consistera à valider la pertinence de mesures composites de l'expérience utilisateur, basées sur des indicateurs de différentes modalités, et en utilisant un protocole de validation multifacette.

ETUDE 3 – MESURE DE L'IMMERSION MULTIMODALE DANS LE CADRE D'APPLICATION DE VIDEOCONFERENCE

Introduction

Parmi les axes de recherche relativement importants au sein des *Bells Labs* figure la communication immersive. Ce sujet fut mis en œuvre par le projet *SlideWorld*,qui avait pour objectif de développer et d'agréger diverses technologies afin de développer des applications de vidéoconférence de nouvelle génération. L'ensemble de ces technologies développées au sein des *Bell Labs* visaient un objectif commun : l'augmentation de la satisfaction de tous les profils utilisateurs, qu'il s'agisse des orateurs, de l'audience locale ou de l'audience distante, en répondant à plusieurs de leurs besoins.

L'un des mots clés de ce projet était l'*immersion*. La perception de ce concept par les équipes de développement se rapprochait de l'état d'absorption cognitive, c'est-à-dire d'une « *disposition à vivre des expériences d'engagement profond, un sens élevé de la réalité de l'objet de l'attention, une imperméabilité aux évènements normalement source de distraction, et une appréhension de l'information par des voies idiosyncrasiques* » (Roche & McConkey, 1990). On peut retrouver dans cette notion de l'immersion des éléments similaires à l'expérience de *Flow* (Csikszentmihalyi, 1990). Certains éléments de cette théorie ont été identifiés comme particulièrement importants dans le cadre des IHM (Trevino & Webster, 1992), tel que le sentiment de contrôle perçu lors de l'interaction avec la technologie, une concentration de l'attention, une élévation de l'excitation, de la curiosité sensitive et cognitive, et une motivation intrinsèque et autotélique de la part du sujet.

Le questionnement du projet SlideWorld était orienté vers l'amélioration de cette immersion via l'identification de fonctionnalités permettant son optimisation, en répondant à certaines questions clefs :

- Comment enrichir et améliorer ce que fournit le système, d'une part à l'orateur, et d'autre part aux auditeurs, dans le fond comme dans la forme ?
- Quels sont les points clés à développer afin d'améliorer l'expérience utilisateur ?
- Comment optimiser l'attention de l'audience et l'interaction entre cette dernière et l'orateur ?

Pour captiver leurs audiences, les équipes travaillant sur *SlideWorld* ont ainsi emprunté des techniques éprouvées du divertissement audiovisuel, par le biais de montages dynamiques obtenus par l'action de multiples caméras munies de fonctions de plan et de zoom contrôlables à distance, le tout orchestré par un algorithme intelligent baptisé « *Virtual Director* », chargé d'effectuer le montage de l'ensemble des plans filmés en temps réel, selon des règles contextuelles. L'idée ici était de rompre avec la monotonie usuelle des présentations filmées dans le cadre professionnel, généralement constituées de larges plans fixes d'un orateur, énonçant son exposé devant une rétroprojection de ses *slides*. D'autres technologies visaient à améliorer le feedback de l'orateur, tel que le « *Jazz Analyser »,* un module d'analyse des

prosodies de l'orateur ayant pour but d'avertir lors d'un ton monocorde, ou encore, le « *Face Detector »,* un module de détection visuelle des expressions du visage, tels que des sourires, afin de sonder en temps réél l'audience distante pour en informer l'orateur.

Toutes ces technologies furent mises au point afin d'améliorer l'expérience utilisateur des utilisateurs, en visant à rendre les vidéoconférences accrocheuses, captivantes et intéressantes. C'est dans ce cadre que la mesure de l'immersion fut identifiée comme outil méthodologique puissant pour le développement de ces technologies, car permettant à l'équipe de discerner les fonctionnalités peu efficaces, de celles qui améliorent réellement le pouvoir de captation des auditeurs. La mesure de l'immersion, comme tous autres construits affiliés à l'expérience utilisateur, est un indicateur mis en avant actuellement par de nombreux auteurs mais difficile à cerner. En effet, nous avons vu dans le domaine de l'évaluation, rares sont les études qui s'appuient sur une méthodologie solide visant l'élaboration d'une mesure robuste de l'expérience utilisateur. Néanmoins, nous avons vu que l'ergonomie a les moyens d'y parvenir, en s'appuyant sur les outils actuellement à sa portée (statistiques, techniques, et informatiques) et le sérieux de ses méthodologies de modélisation des construits complexes. Nous pensons que pour cela, il convient de s'appuyer sur une phase de conceptualisation solide, une opérationnalisation multi-mesures et une phase de validation adaptée. De plus, dans le contexte de la mesure d'un construit abstrait comme l'UX, nous pensons qu'il convient de s'appuyer sur une approche complémentaire, celle de la triangulation multimodale, en s'appuyant sur un large spectre d'indicateurs, d'ordre physiologiques, comportementaux et auto-rapportés. Ces mesures composites seront testées à partir de la théorie de la généralisabilité, dont le modèle multifacette est justifié par la complexité des facteurs à prendre en compte dans le domaine de l'évaluation des applications de communication. Cette thèse se situant dans un cadre industriel, cette étude conduira à l'établissement de protocoles d'évaluations visant un rapport optimal entre la qualité et l'effort de mesure.

La conceptualisation de l'immersion

Il est de plus en plus courant, dans les modèles actuels de l'expérience utilisateur (Mahlke, 2008; Napoli, 2010), de trouver en place des choix des construits tels que le flow d'immersion, d'absorption ou d'engagement. Ces construits de « fin de chaine », synthétiques par nature, sont particulièrement intéressants pour la mise au point d'une mesure globale de l'UX. En effet, résultant de l'impact d'une série de qualités UX en amont, ils sont moins soumis à la contingence. Néanmoins, à cause de nombreux praticiens qui pensent que « *nommer* » est équivalent à « *définir* » (MacKenzie et al., 2011), nous avons vu fleurir de nouveaux construits sans savoir en quoi ils différaient de ceux qui existaient déjà. Ainsi, il existe actuellement une multitude de concepts de haut niveau se rapportant de près ou de loin à la notion d'engagement, tels que :

- le « *flow* » (Cowley et al., 2008; Hektner et al., 2007),
- l'immersion (Jennett et al., 2008),
- la présence (Wijnand Ijsselsteijn & Riva, 2003; Giuseppe Riva et al., 2004),
- l'absorption cognitive (Agarwal & Karahanna, 2000; Saadé & Bahli, 2005),

- la jouissance (« enjoyment » ; Vorderer, Klimmt, & Ritterfeld, 2004; Weber, Tamborini, Westcott-Baker, & Kantor, 2009),
- l'engagement (O'Brien & Toms, 2008; Jane Webster & Ahuja, 2006),
- l'implication (W. Lee et al., 2011; Tung & Deng, 2006),
- l'intérêt (Thiran et al., 2010),
- ou encore, l'expérience ludique (Mirza-Babaei & McAllister, 2010; Poels et al., 2008).

Il est difficile de discerner conceptuellement toutes ces entités, parfois nommées différemment mais associées au même concept, et parfois nommées de manière identique mais renvoyant à des concepts différents. Par exemple, l'immersion peut être comprise parfois dans le sens de l'illusion de présence physique (Wijnand Ijsselsteijn & Riva, 2003) et parfois comme le fait d'être absorbée mentalement dans une activité (Agarwal & Karahanna, 2000). La délimitation des construits est importante car il est impossible de mesurer correctement ce que l'on a du mal à cerner conceptuellement. Il conviendra donc d'identifier précisément ce que l'on souhaite mesurer, en discernant précisément les caractéristiques propres au construit et les différences qui le séparent des notions approchantes. C'est ce que nous tenterons de faire avec l'immersion car cela constitue la première étape vers la création d'une mesure robuste.

La notion d'immersion

Dans la littérature, il est fréquent de parler d'immersion pour décrire les moments les plus intenses de l'expérience. Par exemple, Holt (1995) signale que l'évaluation de l'expérience se fait généralement a posteriori ou en dehors des moments d'immersion. Cela veut dire que, lors de l'immersion, l'individu n'a ni le temps ni la disponibilité d'entamer une évaluation de son vécu. La jouissance (« *enjoyment* » ; Vorderer, Klimmt, & Ritterfeld, 2004; Weber, Tamborini, Westcott-Baker, & Kantor, 2009), l'implication (Lee et al., 2011; Tung & Deng, 2006) ou l'intérêt (Thiran et al., 2010) sont, eux, plutôt des conséquences de l'immersion, sans pour autant être toujours activées lors d'une expérience captivante (qui peut susciter une émotion forte sans être forcément chargée positivement). L'expérience ludique (Mirza-Babaei & McAllister, 2010; Poels et al., 2008) est une notion plus large, qui voit l'immersion comme un facteur critique et une conséquence d'une bonne expérience de jeu (C Jennett et al., 2008). Certains concepts, propre à des domaines particuliers, proposent des notions entrant pleinement dans le champ de l'immersion, sans en utiliser le terme. Par exemple, dans le domaine de l'éducation, les auteurs parleront plutôt d'engagement. Pour eux, il se manifeste sous la forme d'attention, d'intérêt intrinsèque, de curiosité et de motivation (Chapman, 1997). O'Brien et Toms (2008) en proposent un modèle (figures 97), montrant à quel point ce concept est proche de l'immersion. Dans le domaine de l'écriture, West, Huber et Sam Min (2004) définissent le phénomène de la transportation narrative comme « *l'immersion dans un texte qui se produit quand sa lecture amène à un changement favorable d'attitude à travers la réduction des réponses cognitives négatives, le réalisme de l'expérience et un transfert affectif* ». D'autres usages du concept d'immersion, au contraire, réduisent son champ à la sensation d'immersion « physique » (Slater & Wilbur, 1997; van den Hoogen, IJsselsteijn, & de Kort, 2009; Visch, Tan, & Molenaar, 2010). C'est pourquoi Lombard et Ditton (1997) distinguent l' « *immersion perceptuelle* », où le système perceptuel de l'utilisateur est complètement stimulé par le monde virtuel et non par le monde physique, de l' « *immersion*

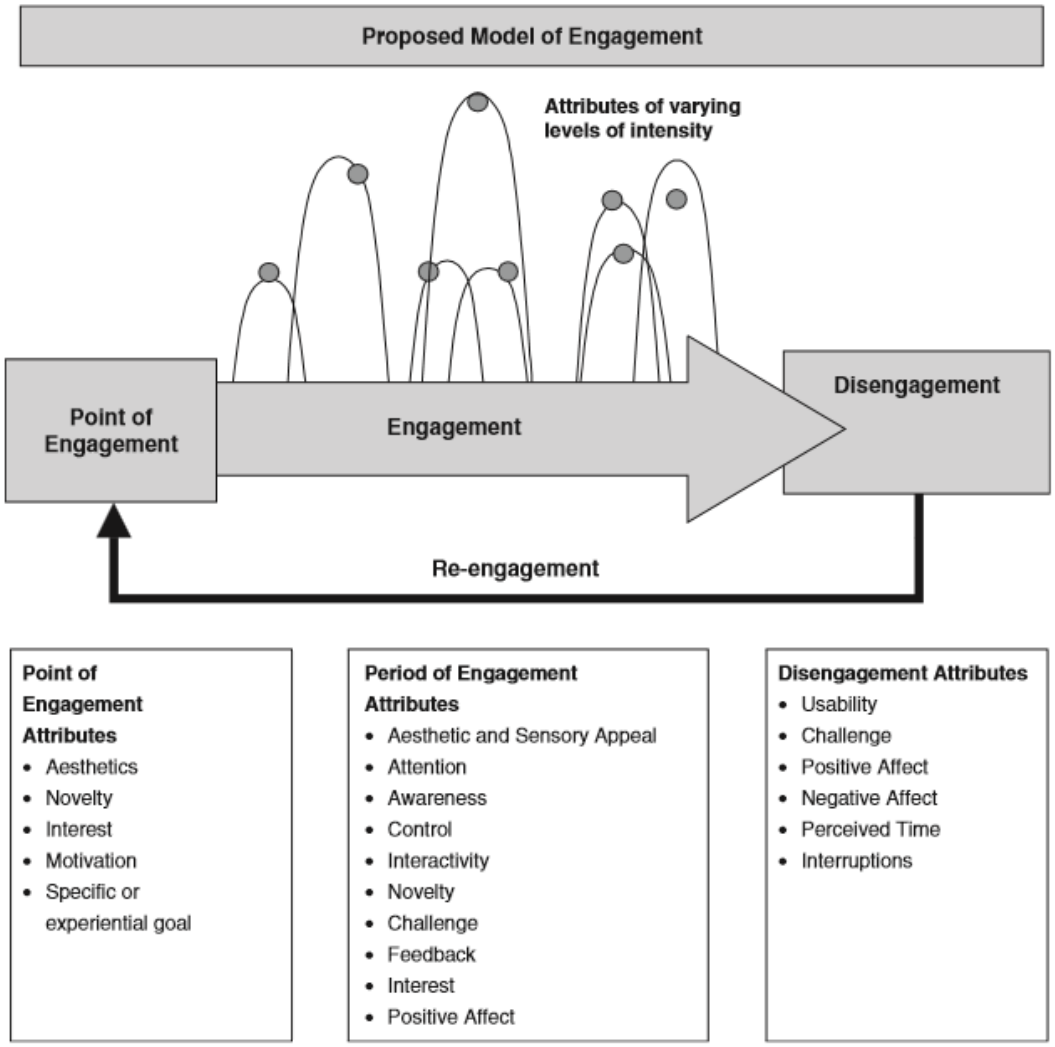


Figure 97 − Modèle de l'engagement et de ses attributs (O'Brien & Toms, 2008)

psychologique », qui décrit un état où l'utilisateur est impliqué, absorbé et totalement engagé. Dans le domaine du jeu vidéo, le modèle de Ermi et Mäyrä (2005) distingue trois types d'immersions : l'immersion sensorielle, qui renvoie au degré de réalisme perceptif et à la crédibilité d'une atmosphère virtuelle ; l'immersion imaginative, permettant au joueur d'être absorbé par la narration ou à s'identifier à un personnage ; et l'immersion basée sur le challenge, dimension plus cognitive et proche de la notion de flow.

Carù et Cova (2003) relèvent certaines faiblesses autour de la conceptualisation de l'immersion. En particulier, ils signalent que la littérature confond le processus d'accès à l'expérience et l'état final qui en résulte. Dans l'étude du processus d'accès à l'expérience, ils proposent un modèle d'appropriation, qui décrit le processus psychologique par lequel le consommateur crée des relations d'attachement durant l'expérience. Ce processus d'appropriation comporte trois étapes : la nidification (faire son nid, création d'un chez soi), l'exploration (découverte et extension du domaine) et le marquage (impression d'un sens particulier). A partir d'une analyse textuelle des récits des personnes ayant assisté à un concert de musique classique, ils mettent en évidence que l'immersion est créée par une combinaison complexe de ces trois phases, au cours desquelles le spectateur met en œuvre ses compétences et ses connaissances. Ils proposent de conceptualiser l'immersion comme un moment fort vécu par le spectateur et résultant d'un

processus partiel ou complet d'appropriation de sa part (Carù & Cova, 2003). L'expérience décrite par les répondants est vécue comme une succession de moments forts, entrecoupés par des moments moins intenses. « *Ce n'est donc pas un plongeon unique, instantané et total* » (Carù & Cova, 2003, p.61). Avec une approche similaire, Brown et Cairns (2004) ont mené une étude qualitative dans laquelle ils ont interviewé de nombreux joueurs sur leurs expérience de jeu. Ils ont constaté que l'immersion était décrite comme le degré d'implication dans un jeu vidéo. 3 étapes ont également été identifiées. La première étape, l' « *engagement* », a lieu quand le joueur découvre les commandes du jeu et se familiarise avec son fonctionnement. Puis, quand les contrôles deviennent « invisibles », la conscience de soi et de l'environnement s'efface peu à peu : « *un état zen où nos mains semble savoir quoi faire, et que notre mental se laisse porter par la narration* » (Brown & Cairns, 2004). C'est la phase d' « *absorption* » (« *engrossment* »). Enfin, le plus haut niveau est appelé « immersion totale » et est décrit par les joueurs comme un sentiment de présence, d'une coupure avec la réalité et vécu comme si seul le jeu comptait à ce moment : « *lorsque vous arrêtez de penser que vous jouez à un jeu et que vous êtes juste dans ce jeu* ». Ce dernier niveau nécessite un niveau très élevé d'attention et représente une expérience de jeu rare et plutôt éphémère, alors que l' « *engagemen*t » et l'« *absorption* » sont plus susceptibles de se produire.

Le flow et l'absorption cognitive

Dans le domaine de la psychologie, Csikszentmihalyi (1997) a introduit la notion de « *flow* » qui présente des similarités avec la notion d'immersion ; une distinction entre ces deux concepts est donc nécessaire. Le *flow* ou « *expérience optimale* » a été analysée dans un large spectre d'activités telles que le sport, le travail, le shopping, les jeux, les loisirs et le numérique. Dans le contexte particulier de la navigation sur internet Novak, Hoffman et Yung (2000) décrivent le *flow* comme un état cognitif vécu par des individus très impliqués dans la navigation sur le Web. Pour décrire l'immersion, ils donnent l'exemple de l'athlète jouant d'une manière exceptionnelle et qui arrive à un état mental au point que rien d'autre ne compte : il est donc complètement immergé. Dans cet état, la personne est entièrement captivée, le temps semble s'estomper et plus rien d'autre ne compte. Le flow est donc une situation d'immersion totale. Néanmoins, la littérature sur le sujet affirme qu'elle correspond à un type particulier d'expérience, c'est-à-dire aux situations où il existe une séquence structurée d'activités et avec un ensemble de buts demandant des réponses appropriées (Csikszentmihalyi, 1997). Le *flow* s'installe quand il existe une convergence entre le niveau d'enjeu et le niveau de compétences requises. Il produit alors une sensation de bien-être provenant du contrôle de la situation. Certaines activités sont plus propices à produire le *flow* (le ski, la navigation sur internet, etc.) que d'autres (regarder la télévision, regarder un film comique, manger au restaurant). Tout comme l'immersion, on mélange souvent les conditions qui amènent au *flow* (le processus qui le caractérise) et ses conséquences (l'état). Ainsi, à partir des composantes du flow proposées par Csikszentmihalyi (1997), Novak, Hoffman et Yung (2000) proposent de considérer, d'une part, la concentration et le contrôle comme les caractéristiques du flow, et d'autre part, la perte de la conscience de soi, la transformation de la perception du temps et le fait que l'expérience devient autotélique[91], comme ses conséquences. Elle partage avec l'immersion la concentration,

[91] Se dit d'une activité qui n'a pour but que l'intense satisfaction qu'elle procure

qui est une de ces caractéristiques fondamentales. En revanche, le contrôle ressenti dans une expérience de flow est particulier à ce type d'expérience, conséquence de la forte cohérence entre les enjeux et les capacités utilisées pour mener à bien l'expérience. Nous pouvons donc conclure que le flow est un cas particulier d'immersion qui émerge dans les activités à fort enjeu, souvent sur le plan cognitif. L'accès à l'expérience se fait par les compétences requises pour le déroulement des activités prévues.

L'expérience extraordinaire (Arnoul et Price, 1993) correspond aussi à une expérience de *flow*, mais d'autres composantes s'ajoutent, telles que le fort contenu émotionnel et l'interaction sociale. En fait, toutes ces expériences possèdent des caractéristiques particulières qui leur confèrent leurs spécificités. Mais une situation ordinaire, c'est-à-dire dépourvue du caractère inhabituel, des forts enjeux ou des forts contenus émotionnels, peut aussi comporter des moments d'immersion totale, comme par exemple la lecture d'un roman. Lors de cette expérience, des moments de concentration totale (cognitive et affective) peuvent arriver : aucun sens autre que la vue n'est stimulée, la personne est coupée de son environnement physique, avec une altération du temps et la perte de conscience de soi. En revanche, aucune sensation de contrôle n'est ressentie, car il n'existe pas d'enjeu important, ni d'interaction. Ainsi, nous constatons que le modèle du *flow* est axé sur une analyse principalement cognitive de l'état d'immersion. Or cet état peut toucher n'importe quelles autres sphères. Ainsi, Fornerino, Helme-Guizon, Gotteland (2006) ont identifiés cinq composantes de l'immersion lors de l'expérience d'un spectacle musicale:

- **Composante Cognitive** : état de haute attention et d'oubli du monde externe
- **Composante Affective** : sensation d'ivresse et d'émotion intense
- **Composante Physique** : réactions corporelles incontrôlables
- **Composante Sociale** : lien renforcé avec les autres et envie de partager l'expérience
- **Composante sensorielle/perceptuelle** : sens en éveil et stimulés

Une notion proche est avancée par Agarwal et Karahanna (2000), qui proposent le concept d'absorption cognitive. Les auteurs décrivent cet état comme une profonde implication dans un logiciel à travers les dimensions de dissociation temporelle, d'immersion focalisé, de plaisir intense, de contrôle et de curiosité. Les antécédents causaux de cet état mental sont l'utilité perçue et la facilité d'utilisation perçue du système, des exigences moindres que pour le *flow*. Le point clé est que l'absorption cognitive invoque de nombreuses expériences similaires au flow (mais en supposant une intensité moindre) sans qu'il soit une expérience optimale. Enfin, nous pouvons également citer l'expérience esthétique, qui partage des éléments similaire avec l'immersion (Michaud, 2012) : (a) perte de contact avec la réalité immédiate avec la sensation d'être transporté dans un autre monde ayant sa propre réalité ; (b) plaisir gratuit (sans utilité propre), parfois confus ou complexe, et qui peut contenir des dimensions cognitives, érotiques, conceptuelles, ludiques, etc.

Synthèse

Nous voyons donc que la notion d'immersion se retrouve en partie ou en totalité dans de nombreux concepts qui ont été avancés dans la littérature. Il convient donc d'en extraire le noyau dur sur lequel nous nous appuieront dans notre étude. Nous essayerons donc de

synthétiser le concept d'immersion à partir de la définition suivante : « *l'immersion est un état final (vs. processus) intense dans lequel l'utilisateur se trouve quand il accède pleinement à l'expérience. Elle peut comporter plusieurs types de manifestations : cognitive, émotionnelle (affective), comportementale, etc.* ». Les deux dimensions fondamentales de cet état d'immersion sont :

- **Attentionnelle** : (i) une concentration cognitive et affective sur le thème de l'expérience, à travers une ou plusieurs manifestations (les pensées et/ou les émotions sont en étroite relation avec l'expérience) ; (ii) le système sensoriel du consommateur est stimulé exclusivement par l'environnement relatif à l'expérience. En conséquence, le consommateur oublie les événements et les objets extérieurs à l'expérience, la perception du temps est modifiée, et seul le moment vécu a de l'importance
- **Motivationnelle** : L'activité produit du plaisir et stimule pleinement l'acteur. Cela le pousse à la continuer sans autre but qu'elle-même (activité autotélique)

La mesure de l'immersion

Il existe deux façons de déceler l'état d'immersion d'une personne. La première façon est d'observer les conditions antérieures donnant lieu à un état immersif. C'est cette méthode qui a été utilisée dans une grande partie des premières études sur le flow, à partir de la mesure conjointe du niveau de difficulté de l'activité et du niveau de compétence de l'individu (Moneta, 2012). La limite majeure de cette approche est que le nombre important de facteurs causaux, leurs interactions et les effets particuliers du contexte rend difficile l'inférence de l'état subjectif d'un utilisateur à partir de ces indicateurs indirects. La deuxième façon d'inférer l'état d'immersion, plus directe, est de se concentrer sur ces manifestations. Ces dernières peuvent être comportementales, physiologiques ou issues de l'auto-évaluation.

L'auto-évaluation

La méthode de recueil la plus simple pour détecter la présence et quantifier l'intensité d'un état d'immersion est de questionner directement celui qui l'expérimente. Les méthodes d'auto-évaluation sont simples, rapides, flexibles, peu coûteuses et fournissent souvent des informations qui seraient difficiles ou impossibles à obtenir par d'autres moyens (Lucas & Baird, 2006). Néanmoins, on constate parfois des divergences entre ce que les personnes rapportent et ce que leur activité psychophysiologique indique. Cela s'explique en partie par la difficulté de s'exprimer sur un état subjectif passé (Stone & Litcher-kelly, 2006) et des connaissances préalables nécessaires sur ce que cela signifie d'être immergé (Ijsselsteijn, Ridder, Freeman, & Avons, 2000). De nombreux auteurs proposent des questionnaires évaluant l'immersion des utilisateurs (Agarwal & Karahanna, 2000 ; Jackson & Eklund ; Novak & Hoffman, 1997; Poels et al. 2006 ; Fornerino, Helme-Guizon, & Gotteland, 2008). Néanmoins, si ces questionnaires d'immersion déclarent mesurer le même concept, de nombreuses différences font qu'il est difficile de considérer ces mesures comme égales entre elles.

La première différence concerne la qualité de construction et de validation des questionnaires. En effet, certains questionnaires abordent l'immersion à travers quelques items construits intuitivement et sans procédure de validation. Par exemple, le modèle PLAY 2009 (H. Desurvire & Wiberg, 2009) contient une heuristique concernant l'immersion (« *Le jeu se sert*

de contenus audio, vidéo et sensitif pour favoriser l'immersion du joueur dans le jeu »), sans que l'on sache à aucun moment ce qu'il faut comprendre par « immersion ». Au contraire, le questionnaire de Flow (Csikszentmihalyi & Csikszentmihalyi, 1988, p. 195) décrit précisément ce qu'il entend par « flow » avant de demander au sujet s'il l'expérimente : « (i) *Mon esprit n'erre pas. Je ne pense pas à autre chose. Je suis totalement impliqué dans ce que je fais. Mon corps se sent bien. Je ne semble pas être distrait par quoi que ce soit. Le monde semble être hors de moi. Je suis moins conscient de moi-même et mes problèmes. (ii) Ma concentration est comme respirer, je n'ai pas à y penser. Quand je commence, le monde qui m'entoure n'existe plus. Je pense que le téléphone pourrait sonner ou la maison pourrait brûler ou quelque chose comme ça. Une fois que je m'arrête je peux laisser le monde qui m'entoure revenir à nouveau. Je suis tellement impliqué que je ne me vois pas comme distinct de ce que je fais* ». Ainsi, le niveau de détails pour décrire le phénomène à évaluer peut varier d'un questionnaire à l'autre, de l'absence totale de description (on se repose sur ce que le mot inspire communément), à une description précise en plusieurs lignes. En complément, de nombreux questionnaires modernes de l'immersion utilisent une abondance d'items pour trianguler ses nombreuses facettes et utilisent des procédures de validation psychométrique avancée (Engeser & Rheinberg, 2008; Moneta, 2012; Qin, Patrick Rau, & Salvendy, 2009). Néanmoins, nous savons que la majorité des questionnaires subjectifs continuent d'être créés *ex nilo* pour les besoins d'une seule étude (Avila et Hornbæk, 2011).

La deuxième différence entre les questionnaire d'immersion se situe en terme de découpage conceptuel. Par exemple, la conceptualisation du flow illustrée par la description fine de Csikszentmihalyi & Csikszentmihalyi (1988, p. 195) correspond à deux construits distincts pour Moneta (2012) :

- **le flow peu profond :** « *Mon esprit n'erre pas. Je ne pense pas à autre chose. Je suis totalement impliqué dans ce que je fais. Mon corps se sent bien. Le monde semble être hors de moi. Je suis moins conscient de moi-même et mes problèmes. Ma concentration est comme respirer, je n'ai pas à y penser. Quand je commence, le monde qui m'entoure n'existe plus. Je suis tellement impliqué que je ne me vois pas comme distinct de ce que je fais »*
- **le flow profond** : « *Je ne semble pas être distrait par quoi que ce soit. Je pense que le téléphone pourrait sonner ou la maison pourrait brûler ou quelque chose comme ça. Une fois que je m'arrête je peux le laisser le monde qui m'entoure revenir à nouveau. »*

D'autres approches multidimensionnelles du flow ont également été proposées au fil du temps, ajoutant à la confusion des modèles de mesure existants. Le modèle le plus connu est composé de neuf dimensions (Csikszentmihalyi, 1990) : (i) Equilibre entre défi et habilité, (ii) Concentration sur la tâche, (iii) Cible claire, (iv) feedback clair et précis, (v) Absence de distraction, (vi) Contrôle de l'action, (vii) dilatation de l'ego, (viii) Altération de la perception du temps et (ix) Expérience autotélique. Certain auteurs (cf. Moneta, 2012) ont proposé de diviser cette structure en deux parties, pour séparer les antécédents du flow (les facteurs qui causent le flow : (i), (iii), (iv) et (v)), des indicateurs du flow (les facteurs qui traduisent les comportements et expériences qui sont causés par le flow : (ii), (vi), (vii), (viii) et (viii)). En complément, de nombreux auteurs ont simplifié le modèle du flow pour ne garder qu'un ensemble de dimensions « clefs », tels que le contrôle, l'attention soutenue, la curiosité et

l'intérêt intrinsèque (Trevino & Webster, 1992; Webster, Trevino, & Ryan, 1993), ou encore, plus simplement, la concentration et le plaisir (Ghani & Deshpande, 1994; Ghani, Supnick, & Rooney, 1991).

Cette disparité en termes de formalisation et de découpage conceptuelle a été constatée également à propos de la mesure de l'absorption cognitive, un concept apparenté au flow. En effet, Agarwal et Karahanna (2000) construisent initialement ce concept autour de cinq dimensions : la dissociation temporelle, l'attention soutenue, le plaisir, le contrôle et la curiosité. Plus tard, Saade et Bahli (2005) réutilisent le concept d'absorption cognitive mais en ne gardant plus que les dimensions les plus centrales à leurs yeux, c'est à dire la dissociation temporelle, l'attention soutenue, le plaisir. De manière similaire, Wakefield et Whitten (2006) conceptualisent l'absorption cognitive sans la dimension du plaisir, car, pour eux, « *combiner le plaisir avec l'échelle d'absorption cognitive masque la variance unique de ce construit affectif* » (p. 294). Cette multiplication des variantes affecte également le domaine de l'immersion dans les jeux vidéo. Par exemple, Brockmyer et al. (2009) développèrent le GEQ (« *Game Engagement Questionnaire* ») qui distingue le flow de l'immersion (« je suis vraiment plongé dans le jeu »), alors que le questionnaire *EGame Flow* (Procci et al., 2012) combine deux dimensions du flow(la dilatation de l'ego et l'altération de la perception du temps) en une seule sous échelle d'immersion. De tels problèmes de conceptualisation de l'immersion dans le domaine a amené de nombreux auteurs à demander des éclaircissements préalables sur le concept afin de déboucher sur une échelle de mesure consensuelle de l'immersion (Brockmyer et al., 2009; Procci et al., 2012).

En effet, la mesure de l'immersion par questionnaire souffre de la disparité des échelles et sous-échelles à cause d'une conceptualisation ne trouvant pas encore consensus. Il conviendra donc de s'appuyer sur les dimensions de l'immersion les plus consensuelles et synthétiques afin d'obtenir une mesure plus stable et moins soumise à la contingence. Sur ce point, les échelles de dissociation temporelle, d'attention soutenue et de plaisir accru, tiré de la mesure de l'absorption cognitive (Agarwal & Karahanna, 2000), sont particulièrement intéressantes, car ces dernières font partie des dimensions de l'immersion les plus utilisées dans le domaine.

Les mesures physiologiques

En complément des mesures subjectives, de nombreux auteurs ont essayé de mesurer l'immersion via l'utilisation de mesures physiologiques. En effet, lorsque notre corps nous prépare à nous engager dans diverses activités, le système nerveux et endocrinien commence à fabriquer et à libérer diverses substances chimiques qui fournissent les bases biologiques des états immersif et affectif. Par exemple, l'activité du cœur et des vaisseaux sanguins augmentent avec l'accomplissement de tâches difficiles et les incitations attrayantes. De même, les stimuli nouveaux, émotionnels, menaçant et attirant l'attention, stimulent l'activité électrodermale. Deux indicateurs psychophysiologiques sont particulièrement intéressants à étudier dans le cadre de l'immersion : les indicateurs de (a) l'état d'attention focalisée (à travers notamment de l'effort mental) et (b) l'état de plaisir.

(a) Dans l'expérience de flow, la conscience est limitée à la tâche effectuée, ce qui implique un degré élevé d'attention sélective et de la charge attentionnelle. Il est intéressant de noter que, malgré une charge élevée et physiologiquement mesurée, cette dernière n'est que peu perçue

subjectivement lors d'une expérience de flow (Bruya, 2010). La littérature sur la détection physiologique de l'effort attentionnel peut donc être un indicateur intéressant pour la mise au point d'une mesure de l'immersion. L'effort mental est lié à des changements d'ordre cardio-vasculaire, respiratoire, facial et sudatoire :

- l'activité musculaire faciale augmente dans le corrugator supercilii (CS), le « *muscle de froncement de sourcils* » (B. H. Cohen, Davidson, Senulis, Saron, & Weisman, 1992; Waterink & van Boxtel, 1994) ;
- la respiration est généralement rapide et peu profonde, avec une augmentation du volume par minute (Backs & Seljos, 1994; Veltman & Gaillard, 1998; Wientjes, 1992);
- les mesures cardiovasculaires montrent généralement une augmentation du rythme cardiaque (diminution de la période cardiaque (HP)) et une augmentation de la pression artérielle systolique (BP), avec une variabilité diminuée dans ces mesures (HRV et BPV, respectivement; Berntson, Cacioppo, & Quigley, 1993; Middleton, Sharma, Agouzoul, Sahakian, & Robbins, 1999; Richter, Friedrich, & Gendolla, 2008; Veltman & Gaillard, 1996, 1998);
- et enfin une augmentation de la sudation (Boucsein, 1992).

Ces observations pointent à l'unanimité une activation accrue de la branche sympathique du système nerveux autonome. Toutefois, il convient de noter qu'au cours d'une tâche allouant une grande qualité d'attention et de mémoire de travail, de bonne performances sont associées à une grande variabilité du rythme cardiaque (HRV), liées à l'influence vagal : c'est la composante parasympathique du spectre HRV (la composante haute fréquence, souvent appelée arythmie sinusale respiratoire ou RSA; Hansen, Johnsen, & Thayer, 2003).

(b) La deuxième branche intéressante à étudier dans le cadre de la mesure de l'immersion est celle des indicateurs physiologiques du plaisir et des émotions. Dans de nombreuses études, l'EMG a été employée avec succès pour différencier les états émotionnels (Niklas Ravaja, Saari, Kallinen, & Laarni, 2006; Witvliet & Vrana, 1995). Deux muscles communément observés dans ce contexte sont le CS, mentionnées précédemment, et le zygomaticus major (ZM), le « *muscle du sourire* ». Les affects positifs et négatifs ont des effets réciproques sur l'activité sur CS, de sorte que les affects négatifs augmenteront son activation alors que les affects positifs la diminuent. L'activité du ZM augmente avec les affects positifs (J. T. Larsen, Norris, & Cacioppo, 2003). L'activité dans ces muscles est également affectée par le niveau de stimulation (« *arousal* »): l'activité du CS est au plus haut dans des situations de valence négative et de faible stimulation, tandis que l'activité du ZM la plus élevée lors d'états positifs et de forte stimulation (Witvliet & Vrana, 1995). Dans le contexte des jeux vidéo, les travaux de Kivikangas (2006) montrent que l'activité du CS est inversement proportionnelle au niveau d'expérience de flow. En ce qui concerne les mesures respiratoires, les états de forte stimulation (« *arousal* ») sont généralement associés à une respiration profonde, haletante et avec un débit inspiratoire élevé (Wientjes, 1992). Les mesures cardiovasculaires, aussi, démontrent une activation du système nerveux autonome spécifiques à certains patterns émotionnelles spécifiques : une stimulation émotionnelle élevée est, d'un point de vue physiologique, généralement associée à une diminution de la période cardiaque et une augmentation de la pression artérielle systolique (Ekman, Levenson, & Friesen, 1983; Schwartz, Weinberger, & Singer, 1981; Witvliet & Vrana, 1995).

A partir de l'état de l'art en physiologie, on peut conclure que l'état d'attention focalisé et l'état de plaisir présente à la fois des corrélats physiologiques compatibles (par exemple, l'activation du système nerveux sympathique) et incompatibles (par exemple, la profondeur des inspirations), ce qui nous permet de formuler des hypothèses sur la physiologie du flux. Tout d'abord, une augmentation du l'immersion devrait être associée à: (i) une diminution de la période cardiaque, (ii) une augmentation du débit cardiaque (diminution de HP et augmentation de BP), (iii) une augmentation du rythme respiratoire, (iv) une augmentation de la conductivité de la peau, qui signifie à la fois une plus grande attention et niveau de stimulation (« *arousal* »). Deuxièmement, le flow pourrait être associé à (v) une augmentation de la profondeur des respirations; (vi) une activité accrue dans le ZM, (vii) une baisse d'activité dans le CS, et (vii) une augmentation du RSA. Cette deuxième série d'associations ne se produit pas en relation à une charge mentale élevée, mais pourrait refléter les affects positifs ressentis lors de l'attention sans effort, propre à l'expérience de flow.

On note que l'utilisation de ces indicateurs dans le domaine des IHM et de l'immersion a été fleurissante ces dernières années, notamment pour les mesures de type cardiaque et de conductance de la peau, facile à mettre en œuvre (Cui & Rau, 2012; J. D. Ivory & Magee, 2009; Keller et al., 2011; Kivikangas et al., 2010; Latulipe et al., 2011; R. L. Mandryk et al., 2006; Mirza-Babaei et al., 2012; Money & Agius, 2009; Niklas Ravaja, Saari, Salminen, et al., 2006; Sammler et al., 2007; Siefert et al., 2009; Sundar et al., 2010; Ulle, 2010; R. D. Ward & Marsden, 2003). Ces études ont confirmé certains des corrélats exposés dans la synthèse précédente bien qu'il soit encore très difficile d'estimer clairement l'utilité réelle de ces indicateurs pour mesurer l'immersion de façon fiable et valide. En effet, il est difficile dans ces études d'estimer la sensibilité de ces mesures car les conditions expérimentales « immersives » sont différentes d'une étude à l'autre, ainsi que les questionnaires subjectifs d'immersion utilisés (et servant de base de comparaison). Par exemple, Nacke et al. (2009) montrent que la conductance de la peau est significativement plus élevée dans la condition « flow » que dans la condition « ennui » et « immersion ». La condition « ennui » se base sur une situation présentant peu de difficultés et des actions répétitives, la condition « immersion » se base sur un environnement complexe, exploratoire et varié et la condition « flow » sur un environnement présentant une difficulté graduelle. On note les échelles de « flow » et « d'immersion », utilisées comme témoin dans cette étude ne sont pas significativement corrélés avec les conditions de test « flow » et « immersion », respectivement. Cet exemple est caractéristique de l'état actuel des études sur le sujet : la construction des conditions expérimentales immersives/non immersives et les questionnaires utilisés ne sont pas standardisés, ce qui pose de véritables problèmes pour comparer les résultats d'une étude à l'autre et pour capitaliser sur les résultats existants.

Autres mesures

D'autres types d'indicateurs peuvent compléter l'éventail d'outils existants pour mesurer l'immersion, tels que les traces écrites, les expressions faciales et corporelles, ou encore les fixations oculaires.

L'analyse des traces écrites peuvent nous donner des informations intéressantes sur l'état de l'utilisateur, via une analyse des sentiments (De Choudhury et al., 2012; Grassi et al., 2011; Nahin et al., 2014; Neviarouskaya et al., 2010; Pan et al., 2010; L. Zhang & Yap, 2012), de la charge cognitive (Khawaja et al., 2014) ou encore de l'engagement (D'Mello & Graesser, 2010). Un des avantages majeurs de l'utilisation de cette technique est son caractère non réactif et écologique (Fritsche & Linneweber, 2006), c'est-à-dire que les données recueillies ne sont pas biaisées comme peuvent l'être celles qui sont obtenues dans un contexte artificiel et provoqué (comme dans un test en laboratoire).

La capture des expressions faciales et corporelles peuvent également nous renseigner sur l'état émotionnel (Bailenson et al., 2008; M. M. Gross, Crane, & Fredrickson, 2010), l'engagement attentionnel (Agliati, Mantovani, Realdon, Confalonieri, & Vescovo, 2006; Asteriadis et al., 2009; D'Mello & Graesser, 2010; Dirican & Göktü k, 2012; C. J. Lee et al., 2006) ou le niveau d'intérêt (Mota & Picard, 2003; Yeasin, Bullot, & Sharma, 2006) des utilisateurs. Pour cela, des techniques de capture vidéo et des algorithmes de reconnaissance, se basant sur le langage corporelle, les postures et expressions faciales ont été mises au point, tel que propose le dispositif *Facereader* (den Uyl et al., 2005; Terzis, Moridis, & Economides, 2010; Bieke Zaman & Shrimpton-smith, 2006).

Enfin, d'autres études dans le domaine nous montrent que les fixations oculaires peuvent être une piste de recherche intéressante pour déceler l'état d'immersion d'un individu dans une activité numérique. Par exemple, l'étude de Jennett et al. (2008) montre que le nombre de fixations d'un individu diminue sur une tâche immersive au cours du temps ; alors que, au contraire, sur une tâche non immersive, le nombre de fixations augmente car l'individu serait plus distrait. Ainsi, une augmentation de l'immersion est associée à une diminution du nombre de fixations oculaires au cours du temps (Figure 98).

Ces nouveaux types d'indicateurs sont encore peu utilisés pour mesurer l'immersion mais présente l'avantage d'être peu intrusifs et coûteux. Seules les fixations oculaires utilisent encore

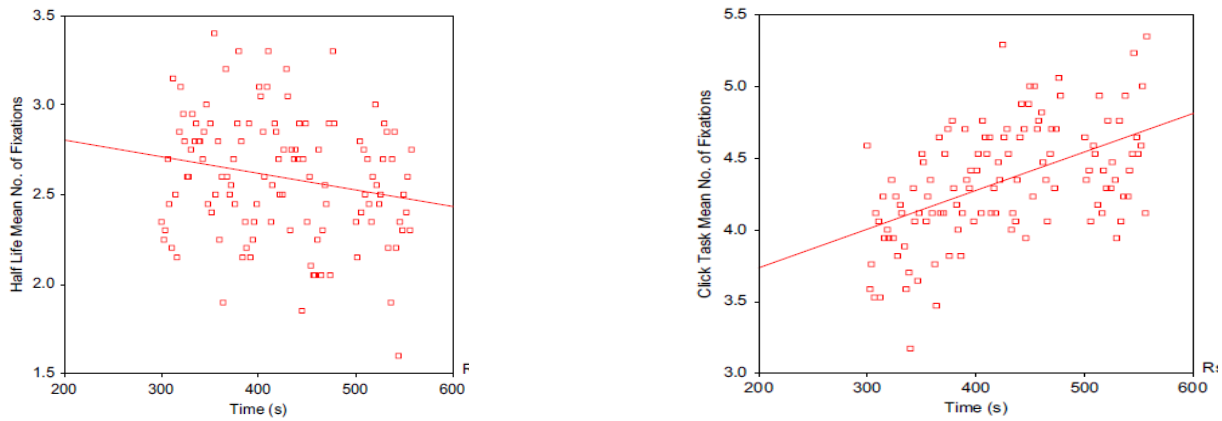


Figure 98 – Moyenne du nombre de fixations par seconde en condition non immersive et en condition immersive, respectivement (Jennett et al., 2008)

un dispositif spécialisé, l'occulomètre, mais des alternatives à partir de webcam classique commence à voir le jour (Sesma, Villanueva, & Cabeza, 2012; Zhang, Bulling, & Gellersen, 2012). Il s'agit donc de pistes intéressantes à explorer pour la mise au point de protocoles d'évaluation de l'immersion triangulant différents types de mesures.

Objectif de l'étude

Le domaine souffre d'une conceptualisation de l'immersion très variable et des protocoles de validation peu stables d'une étude à l'autre. Afin de tester le potentiel de ces mesures sur une base commune, il convient de construire un protocole de tests allant au-delà des méthodes classiques de validation. De plus, en complément de l'étude individuelle des indicateurs de l'immersion, l'analyse psychométrique d'indicateur composite sera réalisée afin de voir si cette piste peut améliorer la mesure de construit complexe comme l'immersion.

Cette troisième étude aura donc pour objectif de tester la validité individuelle d'un ensemble de mesure de l'immersion, puis de tester un certain nombre de mesures multimodales de l'immersion. Il s'agit donc, dans un premier temps, d'identifier les mesures les plus pertinentes, puis, dans un second temps, de tenter de les combiner pour améliorer la qualité de la mesure. En effet, nous pensons que l'utilisation de données de modalités différentes, par complémentarité, peut permettre de contrebalancer les limites respectives de chacune de ces mesures prise en isolation. Pour ce faire, nous utiliseront un protocole de validation multifacettes basé sur la théorie de la généralisabilité, que nous pensons plus adapté que les modèles de validation bi-facettes utilisés communément dans le domaine des IHMs.

L'analyse psychométrique de ces mesures a été réalisée à partir de deux expériences distinctes, en utilisant le cas d'usage de la vidéo-conférence pour manipuler le niveau d'immersion des utilisateurs. La première expérience a été réalisée sur 60 utilisateurs et a permis de recueillir des données de type physiologiques (conductances de la peau), vidéos (expressions faciales), comportementales (verbalisations écrites) et auto-rapportées (questionnaires). Elle a été complétée par une autre expérience sur 90 utilisateurs, ce qui a permis de récupérer d'autres types d'indicateurs (questionnaire, expression faciale, conductance de la peau, rythme cardiaque, comportement oculaire) et de fournir des données suffisantes pour tester la pertinence des protocoles de mesures multimodales dans le domaine particulier de l'immersion. Comme pour l'étude 2, ces indicateurs ont été choisis pour leur relative facilité de recueil (ou en misant sur une amélioration future et probable de celle-ci) et pour leurs grandes complémentarités, afin de posséder une mesure de l'immersion la plus robuste possible.

Dans un premier temps, cette analyse a permis d'identifier les indicateurs les plus pertinents pour mesurer l'immersion, grâce à l'étude de leurs qualités psychométriques respectives : ces résultats sont exposés dans **l'analyse psychométrique mono-mesure**. L'analyse des résultats s'est ensuite poursuivie par la triangulation des indicateurs d'immersion jugés pertinents et l'estimation de leur validité par un protocole de test multifacette : ces résultats sont exposés dans **l'analyse multifacette multi-mesure**.

Analyse mono-mesure (Expérience 1)

Méthodologie

Plan expérimental

Pour tester la fiabilité des mesures, un plan expérimental à deux facteurs (3 x 2), manipulant le « type d'interface » (« *Chat 1/3* », « *Chat 2/3* », « *Sans Chat* ») et le « *Style de présentation* » (« *Bonne présentation* » et « *Mauvaise présentation* ») a été élaboré (Figure 99). Chaque participant a été assigné aléatoirement à une des trois interfaces (condition emboitée) et à chacun des deux styles de présentation, successivement (condition croisée). Pour contrôler l'effet d'ordre, le style de présentation a été contrebalancé entre les participants.

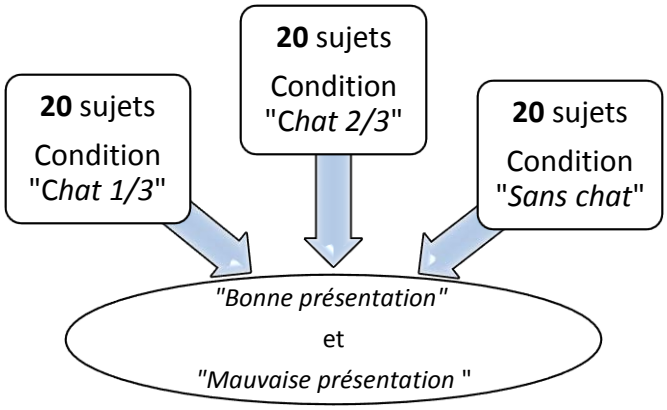


Figure 99 – Plan expérimental et répartition des utilisateurs

Artefact de test

Maquette interactive (AttentionApp)

Une maquette interactive de l'interface *SlideWorld* a été développée pour manipuler le niveau d'immersion des auditeurs distants à partir de différentes interfaces. Un module de chat y a été intégré pour permettre la capture de traces écrites. Ces interfaces se distinguent par la taille et la présence de deux éléments : le « module de Chat » et le cadre vidéo. L'interface « *Chat 1/3* » est composée par un module de chat mesurant un tiers de l'écran et un cadre vidéo comblant les 2/3 restants (Figure 100). Ce type d'interface correspond à l'écran de base de l'interface *SlideWorld*. L'interface « *Chat 2/3* » est composée par un module de chat mesurant 2/3 de l'ecran et un cadre vidéo comblant le tiers restant (Figure 101). L'inversion des proportions du module de chat et du cadre video a pour but de réduire le niveau d'immersion de l'utilisateur en dégradant son expérience par un cadre de présentation très petit et des éléments de distraction plus prégnants. L'interface « *Sans Chat* » est composée seulement d'un cadre vidéo mesurant 2/3 de l'écran (Figure 102). L'absence de chat permet de disposer d'une condition expérimentale sans source de distraction.

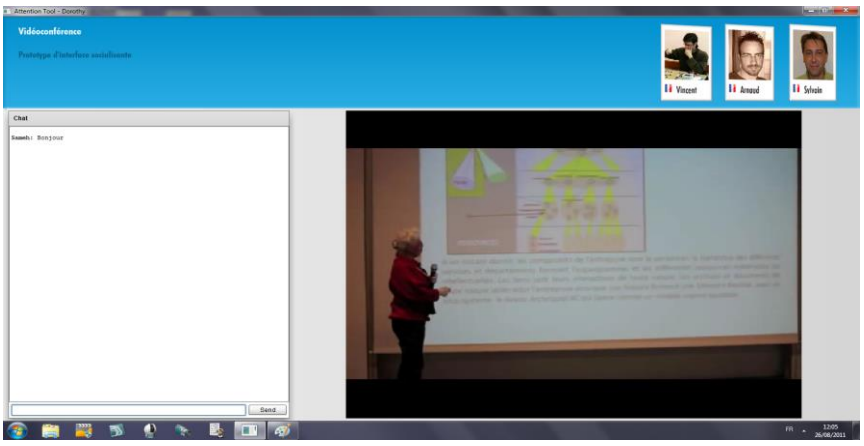


Figure 100 – Capture d'écran de l'artefact de test dans la condition « chat 1/3 »

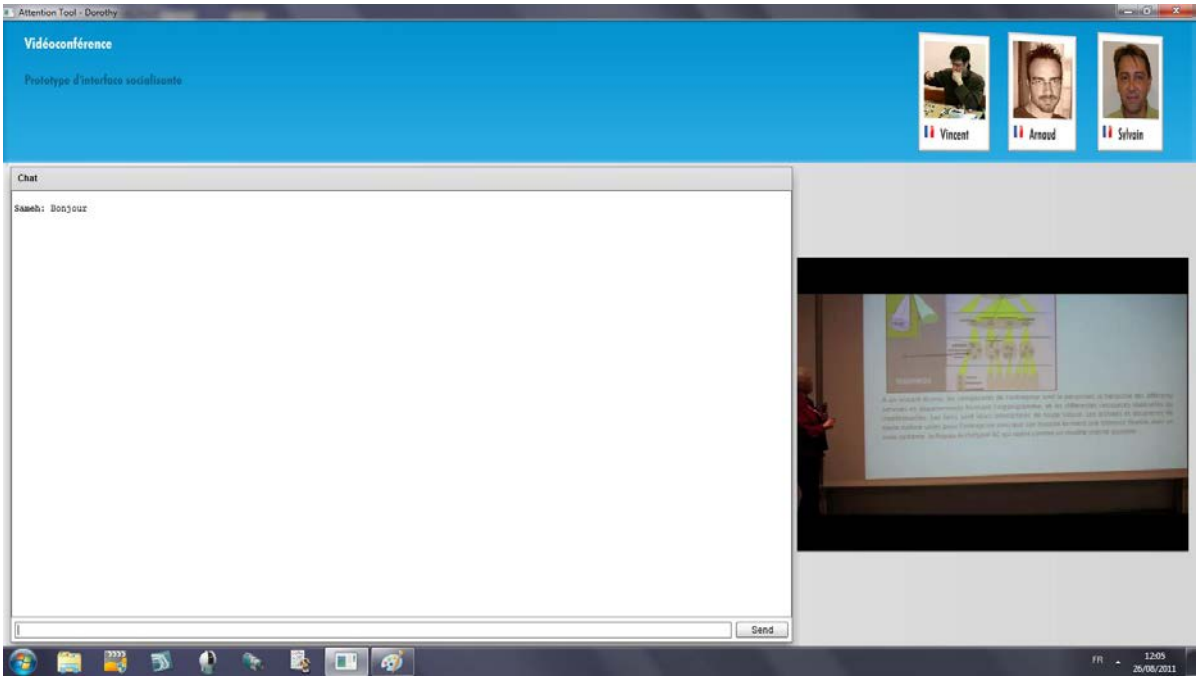


Figure 101 − Capture d'écran de l'artefact de test dans la condition « chat 2/3 »

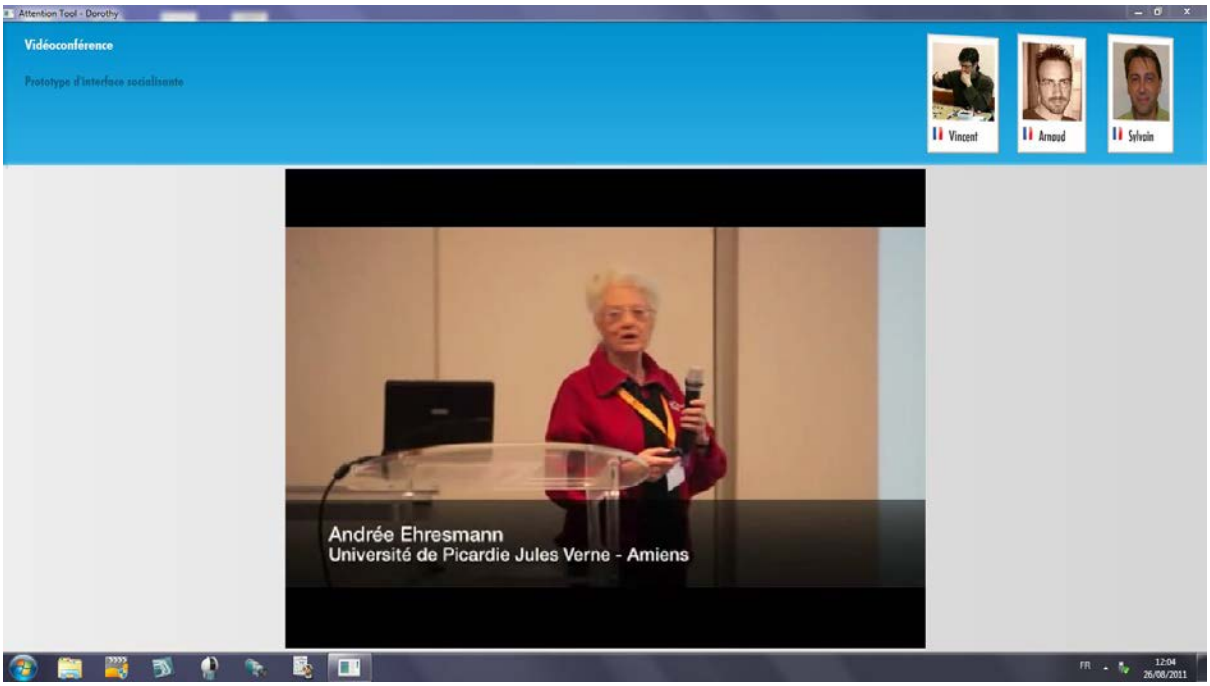


Figure 102 − Capture d'écran de l'artefact de test dans la condition « chat 2/3 »

Interface de contrôle « magicien d'Oz » (AttentionWiz)

Etant requis pour la condition « Chat 2/3 » et « Chat 1/3 », une interface utilisant la technique du magicien d'Oz (Jay Bradley & Benyon, 2009) a été crée, pour simuler la présence de trois voisins virtuels durant l'expérience (Figure 103). Cette interface est localisée sur un ordinateur spécifique, en dehors du champs de vision du participant. Sur cette interface, la fenêtre du haut

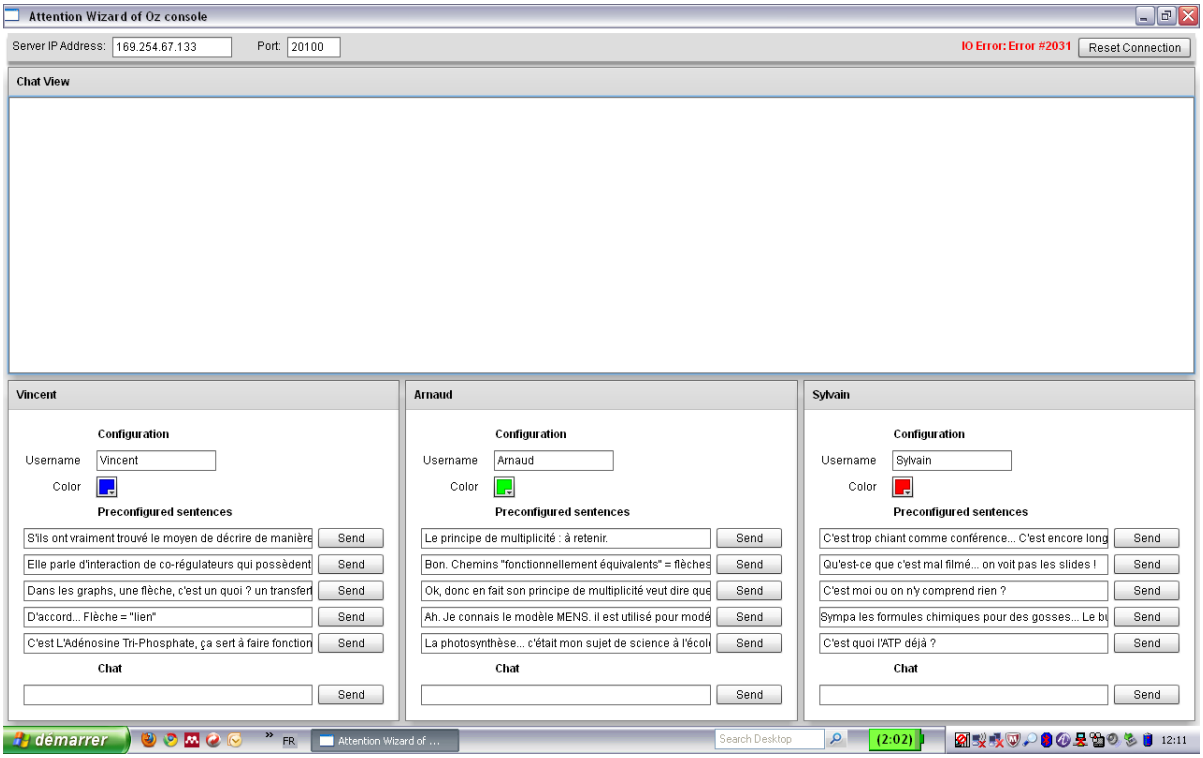


Figure 103 - Capture d'écran de l'interface de contrôle à distance du chat.

affiche la zone de discussion du participant**.** Les fenêtres en dessous correspondent chacune à un des trois voisins virtuels. Le modérateur contrôle ces co-auditeurs virtuels en tranmettant des phrases pré-enregistrées ou de façon manuelle pour simuler des interactions naturelles et plausibles. La personnalité de ces trois auditeurs distants sont basés sur des archétypes prédéfinis :

- **Le *questionneur* :** il pose constamment des questions auxquelles les réponses sont la plupart du temps triviales. Quelques-unes de ces questions sont pertinentes, mais s'il se montre particulièrement perturbateur lors d'un passage défini de la vidéoconférence, où plusieurs informations essentielles sont délivrées par le présentateur.
- **Le *commentateur avisé*** : il connaît bien le sujet abordé par la présentation, reformule certaines informations essentielles en leur apportant quelques précisions qu'il juge pertinentes, à tort comme à raison. Il veille à distiller alternativement des précisions importantes et des détails dispensables.
- **Le *distracteur*** : il est peu attentif au discours de l'orateur, et cherche à provoquer de l'amusement lors de la vidéoconférence en distribuant quelques traits d'humour de temps en temps. Il a tendance à adopter un comportement qui cumule deux caractéristiques à fort potentiel d'interférence, c'est-à-dire bavard et hors de propos.

Chaque voisin virtuel est intervenu au moins cinq fois durant la vidéo conférence. En fonction des réactions des sujets, des relances ont été émises. Elles n'excèdent toutefois pas trois occurrences consécutives par minute, afin de ne pas saturer les sujets avec un nombre excessif de messages.

Présentations vidéos

Figure 104 - Vidéo « Bonne présentation »

Figure 105 - Vidéo « Mauvaise présentation »

Afin de manipuler le niveau d'immersion des participants, deux vidéos de même durée similaires et différentiées par leurs qualités didactiques et pédagogiques, ainsi que par leurs qualités d'édition, ont été choisies. La vidéo « *Bonne présentation* » (7'18'') est extraite d'un programme sur le cycle de la vie. Le style de présentation est didactique, dynamique, bien cadrée et présente une bonne intéraction avec l'audience (Figure 104). La vidéo « mauvaise présentation » (7'43'') est extrait d'une conférence de mathématique appliquée sur la complexité et l'émergence. La présentation se compose d'un long monologue non structuré, mal articulé et mal cadré (Figure 105).

Matériel physique et configuration logicielle

Sur le plan technique trois ordinateurs et un système d'acquisition physiologique BIOPAC (Figure 106) ont été mobilisés pour l'expérience. Un PC *Sony VAIO* a été utilisé comme ordinateur de test : il héberge l'enregistrement de la session via *MORAE* et l'exécution de la maquette de l'interface de *SlideWorld* côté auditeur distant (*AttentionApp*). Un PC *Lenovo Thinkpad* a été utilisé comme poste de modération à partir du module de commande *Wizard of Oz* (A*ttentionWiz)*. Un PC *Asus ROG G73Sw* a servi de poste d'acquisition et de contrôle du signal *GSR* via le logiciel *Acqknowledge*. L'ensemble des postes informatiques et le *BIOPAC MP150* ont été reliés par un *Switch Ethernet 3Com GSU05* et des câbles *RJ45*. Enfin, un thermomètre a permis de contrôler la température de la salle de test avant chaque passation.

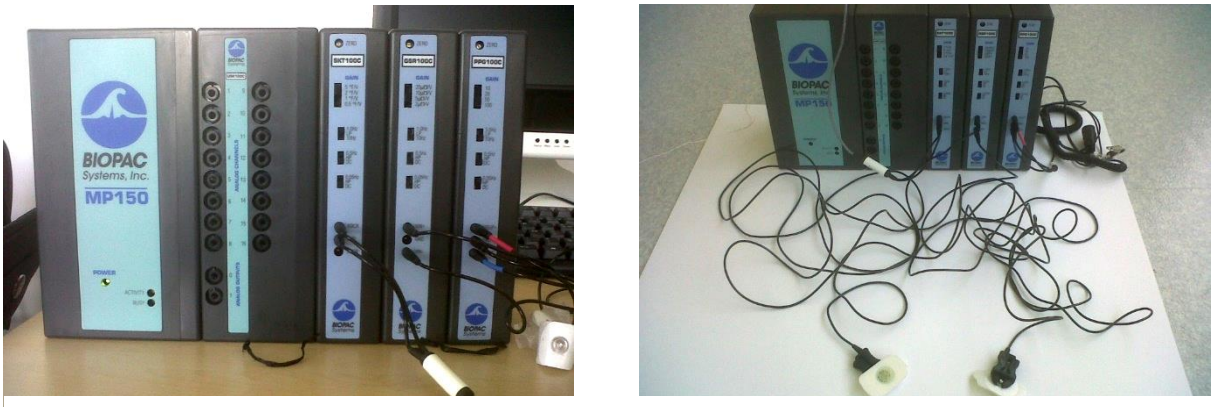


Figure 106 - Système d'acquisition de données BIOPAC MP150, Amplificateur GSR100C et Electrodes EL507»

Mesures

Auto-rapporté (Questionnaire d'immersion)

Un questionnaire immersion (échelle de Likert en 5 points), utilisant des items de trois sous-échelles du questionnaire d'absorption cognitive (Agarwal & Karahanna, 2000) a été mis au point pour les besoins de l'étude. Ce questionnaire reprend 9 items du questionnaire initial, puis ont été adaptés au contexte de l'étude, traduit de l'anglais au français et contrôlé par deux collègues de langue maternelle anglaise (Tableau 18). La première sous-échelle utilisée (3 items) est celle du plaisir accru (« *Heightened Enjoyment* ») et correspond au sentiment de plaisir ressentie lors de l'expérience d'intéraction. Cette dimension de l'immersion est une synthèse de l'intérêt intrinsèque du flow (Webster et al., 1993) et du plaisir perçu (F. D. Davis, Bagozzi, & Warshaw, 1992). Deux autres sous-échelles du questionnaire d'Agarwal & Karahanna (2000) ont été utilisés et se rapporte directement à l'expérience d'absorption cognitive : la dissociation temporelle (2 items) et l'immersion concentrée (4 items). La dissociation temporelle (« *temporal dissociation* ») correspond à notre inaptitude à enregistrer le passage du temps lors d'un engagement profond. L'immersion concentrée (« *focused immersion* ») se rapporte à l'expérience d'engagement totale où toutes les autres demandes attentionnelles sont, par essence, ignorées. Ce questionnaire est rempli par tous les utilisateurs (n = 60) et à la fin de chaque présentation vidéo. De plus, ce questionnaire a été repris à l'identique dans le cadre de l'expérience 2 pour permettre la comparaison des deux expériences et de toutes les mesures de l'immersion utilisées.

Tableau 18 – Questionnaire d'immersion (adapté d'Agarwal & Karahanna, 2000)

Sous-échelles	Items
Plaisir accru	Je me suis amusé en regardant cette vidéoconférence
	J'ai aimé regarder cette vidéoconférence
	Regarder cette vidéoconférence m'a ennuyé
Dissociation temporelle	Le temps semblait s'écouler très vite lorsque je regardais cette vidéoconférence
	Lorsque je regardais cette vidéoconférence, je perdais parfois la notion du temps
Immersion concentrée	Lorsque je regardais cette vidéoconférence, j'étais capable de bloquer la plupart des autres distractions
	Lorsque je regardais cette vidéoconférence, j'étais immergé dans la tâche que j'effectuais
	Lorsque je regardais cette vidéoconférence, je me laissais facilement distraire par d'autres choses qui captaient mon attention
	Lorsque je regardais cette vidéoconférence, mon attention ne se laissait pas distraire très facilement

Conductance de la peau (GSR)

La capture de la conductance de la peau a été réalisée grâce au système d'acquisition de données BIOPAC MP150, d'un Amplificateur du signal galvanique GSR100C et de deux Electrodes EL507. Les électrodes de mesure de la réponse électrodermale ont été placées sur la deuxième phalange de l'index et du majeur, comme l'illustre la position **#1** de la figure 107. Elles ont été placées sur la main la moins dominante du participant et de manière à ce que la frappe au clavier perturbe au minimum l'enregistrement des données physiologiques.

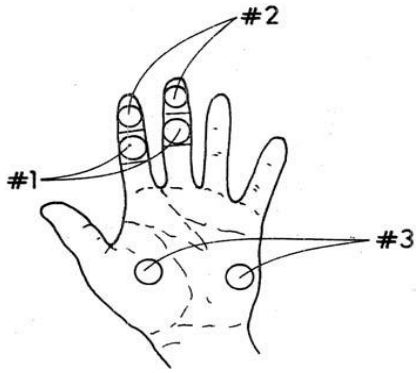


Figure 107 – Trois possibilités de placement de paires d'électrodes EL507 (Cacioppo, Tassinary, & Bernston, 2007)

3 indicateurs ont été recueillis : la conductivité de la peau moyenne (GSR_{MOY}), la Baseline (GSR_{BASE}), et la conductivité de la peau relative ($GSR_{REL.}$). La conductivité de la peau moyenne (en millivolts) a été calculée par la moyenne des données physiologiques obtenue toutes les secondes durant toute la période de vidéoconférence. Cet indicateur brut ne prend pas en compte la variabilité interindividuelle du niveau de base de la réponse galvanique. La Baseline a été obtenue en calculant la moyenne des données physiologiques obtenue durant les 30 dernières secondes de la période de relaxation. Cela a permis de recueillir le niveau de base de la réponse galvanique des participants avant l'expérience de la vidéoconférence. Enfin, la conductivité de la peau relative correspond à la différence entre la conductivité de la peau moyenne et la Baseline (GSR_{MOY} - GSR_{BASE}). Cet indicateur prend en compte la variabilité interindividuelle du niveau de base en nous donnant l'écart relatif de la réponse galvanique par rapport à ce niveau. Ces données ont été capturées pour 45 utilisateurs, ce qui correspond à 15 utilisateurs pour chacune des trois conditions expérimentales concernant l'interface.

Traces écrites et expressions faciales/corporelles

Le recueil des traces écrites et des expressions faciales/corporelles a été réalisé à partir des enregistrements MORAE. Les verbalisations écrites ont ainsi été récoltées par la capture des saisies au clavier provenant du module de chat et les expressions faciales par la Webcam du PC *Sony VAIO* (format vidéo *asf* 320*240).

De plus, une évaluation experte en aveugle (sans connaissance préalable des niveaux d'immersion mesurés et déclarés par les utilisateurs observés) a été réalisée sur un échantillon des données. Ainsi, cinq experts ont procédé à une inspection individuelle de ces vidéos (n=30) et de ces traces écrites (n=60). Ces données ont été anonymisées et présentées dans un ordre aléatoire. Pour chacune d'entre elles, il a été demandé d'évaluer le niveau d'immersion et de plaisir de l'utilisateur en question à partir d'une échelle de Likert en 6 points. De plus, pour chacune de ces estimations, il a été demandé aux évaluateurs de donner un indice de confiance de leur jugement (1 = Incertain, 6 = Certain). Les consignes précises et les grilles de notation sont disponibles en annexe 3A.

Ainsi, pour l'évaluation de l'immersion et du plaisir, nous avons recueillis en tout 8 indicateurs : 4 pour les traces écrites et 4 pour les expressions faciales et corporelles. Pour les traces écrites,

il s'agit de l'évaluation experte de l'immersion et du plaisir, ainsi que leurs indices de confiance respectifs, pour 30 utilisateurs et pour les deux présentations (30*2=60). En effet, l'analyse des traces écrites a été réalisée sur toutes les données disponibles, 15 utilisateurs n'ayant pas eu accès au module de Chat dans la condition « *Sans Chat* ». Pour les expressions faciales et corporelles, il s'agit de l'évaluation experte de l'immersion et du plaisir, ainsi de leurs indices de confiance respectifs, pour 15 utilisateurs et pour les deux présentations (15*2=30). En effet, l'analyse les expressions faciales et corporelles a été réalisée sur 30 vidéos sur les 120, soit approximativement 4 heures d'enregistrement vidéo. Enfin, 4 indicateurs supplémentaires ont été créés en pondérant les évaluations expertes par leurs indices de confiance respectifs.

Recrutement et procédure

Figure 108 – Ordre de réalisation du test

Les participants ont été recrutés au sein des Bell Labs. La langue courante a été contrôlée préalablement au recrutement, car elle est susceptible d'influer sur la compréhension. Par précaution, seules des personnes parlant couramment le français depuis plus de dix ans ont été recrutées. De plus, pour ne pas biaiser les enregistrements de la conductance de la peau, les participants ne doivent pas avoir les mains excessivement fripées, excessivement froides (doigts virant au bleu), et ils ne doivent pas présenter de trouble de la sudation telle que l'hypersudation ou encore la maladie de Raynaud (doigts virant au blanc). De ce fait, ces prérequis ont été systématiquement contrôlés par une observation informelle des mains des sujets lors des contacts de recrutement.

Avant chaque passation, la température ambiante a été systématiquement contrôlée pour vérifier si celle-ci était comprise entre 22°C et 25°C. Au début de la passation, le participant a été assigné aléatoirement à l'une des trois conditions basées sur l'interface : *« Sans Chat* », « *Chat 1/3* » et « *Chat 2/3* ». Les participants ont été ensuite installés, les capteurs GSR ont été branchés et la vérification de l'assise du sujet et de la posture correcte de la main (paume orientée vers le haut) a été réalisée. Puis, le participant a été accompagné pour la lecture des consignes générales et le casque audio a été mis en place. Les participants ont été informés qu'ils allaient devoir regarder deux vidéoconférences. Pour les conditions « Chat 1/3 » et « Chat 2/3 », les participants ont été informés de la possibilité de discuter avec les autres auditeurs, via un module de chat. Ensuite, le participant a bénéficié de cinq minutes de repos pour établir la Baseline de la GSR. Puis, la première vidéo a été diffusée. Pour les conditions « Chat 1/3 » et « Chat 2/3 », l'expérimentateur est intervenu en simulant les différents voisins virtuels par l'interface *Wizard of Oz*. A la fin de la première vidéo, le questionnaire d'immersion a été présenté, suivi par cinq nouvelles minutes de repos pour rétablir la Baseline de la GSR. Enfin, la deuxième vidéo a été diffusée, et la passation a été terminée par la deuxième présentation du questionnaire d'immersion. L'ordre général du test est représenté sur la figure 108. Les consignes ont été directement intégrées à MORAE. Elles sont disponibles en intégralité en annexe 3B.

Résultats

Descriptif de l'échantillon et des conditions de passation

Soixante sujets, 18 femmes et 42 hommes ont participé à l'expérimentation. L'âge moyen des sujets était de 31 ans (σ = 9,02) et l'âge médian de 28 ans. Plus de la moitié des participants (55%) avaient un niveau d'études équivalent à BAC+5, et plus d'un cinquième (21%) d'entre eux un niveau d'études équivalent à BAC+8.

La température ambiante moyenne en début d'expérience était de 24,22°C (σ = .74) et la température ambiante moyenne en fin d'expérience était de 24,36°C (σ = .72), soit un écart de température moyen de +0,14°C (σ = .22). Sur les 45 passations comportant des mesures physiologiques, une variation de la température fut observée pour 17 (37,78%) d'entre elles : dix augmentations de 0.5°C, six augmentations de 0.25°C, et une diminution de 0.25°C. La durée moyenne des passations était de 35 minutes et 30 secondes (σ = 3'51'').

Mesure auto-rapportée (Questionnaire d'immersion)

Les analyses statistiques sur le questionnaire d'immersion ont été réalisées sur les 60 participants de l'expérience. La consistance interne du questionnaire, mesurée par le coefficient alpha de Cronbach, est bonne (α = .852), ainsi que pour la sous-échelle de plaisir accru (α = .852). La consistance interne des sous-échelles de dissociation temporelle (α = .676) et d'immersion concentrée (α = .683) sont moins élevées mais reste correctes. Les corrélations entre les différentes sous-échelles du questionnaire nous montrent des liens forts et très significatifs (p > .001) entre les trois dimensions mesurées (tableau 19).

Tableau 19 – Corrélations entre les trois sous-échelles du questionnaire d'immersion

	Plaisir accru	Immersion concentrée
Immersion concentrée	**.471****	-
Dissociation temporelle	**.572****	**.552****

** p < .01

Une analyse de la variance (Tableau 20), nous montre que le questionnaire d'immersion permet de discriminer significativement entre la « bonne » et la « mauvaise » présentation ($F(1,57)$ = 37,419, p > .000, η^2_p = .396) mais pas entre les différentes interfaces (p < .05). Le questionnaire d'immersion a permis également de détecter un effet d'interaction entre le style de présentation et le type d'interface ($F(2,57)$ = 6,038, p = .004, η^2_p = .175).

Tableau 20 – Analyse de la variance du questionnaire d'immersion en fonction du type d'interface et du style de présentation

Sources	**dll**	**F**	η^2_p
Type d'interface (I)	2	1,641	-
Style de présentation (P)	1	**37,419****	.396
I x P	2	**6,038***	.175
Erreur	57		

*p > .05 **p > .001

La figure 109 illustre cet effet d'interaction entre le style de présentation et le type d'interface. On constate un effet plancher sur l'immersion de lors du visionnage de la mauvaise présentation et un impact positif du Chat 1/3 lors du visionnage de la bonne présentation.

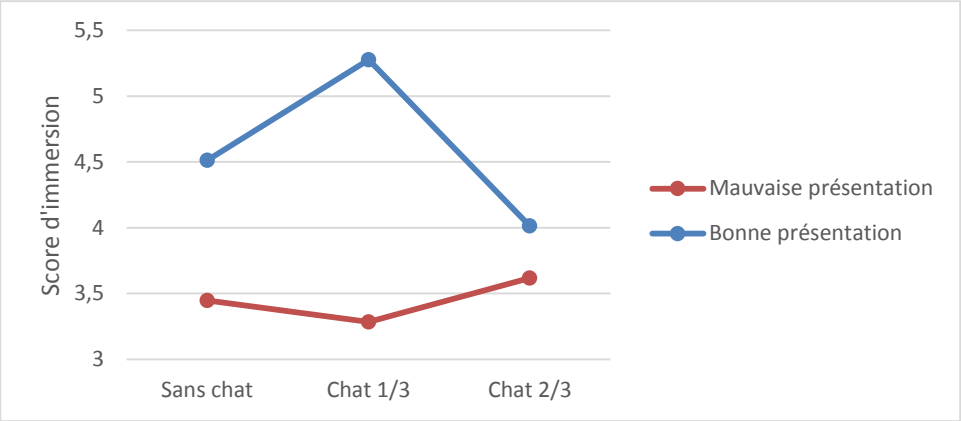


Figure 109 – Score d'immersion en fonction du type d'interface et du style de présentation

Conductance de la peau

Les analyses statistiques sur les données physiologiques de la conductance de la peau ont été réalisées sur les 45 participants ayant été soumis à cet enregistrement. Pour la conductivité moyenne de la peau, aucun lien significatif avec l'échelle auto-rapportée d'immersion et ses sous-échelles n'a été trouvé (Tableau 21). Pour la conductivité de la peau relative, on constate une corrélation négative faible avec la dimension d'immersion concentrée (r = -.362, p < .01) et le score d'immersion total (r = -.290, p < .01).

Tableau 21 – Corrélations entre les indicateurs GSR et les échelles du questionnaire

	Conductivité de la peau moyenne	Conductivité de la peau relative
Plaisir accru	.004 (n.s.)	-.203 (n.s.)
Immersion concentrée	-.200 (n.s.)	**-.362***
Dissociation temporelle	-.099 (n.s.)	-.162 (n.s.)
Score d'immersion total	-.166 (n.s.)	**-.290***

*p < .01

Une analyse de la variance (Tableau 22), nous montre que la conductivité de la peau relative a permis de discriminer significativement entre les types d'interfaces (F(2,42) = 3,553, p = .04) avec une taille d'effet plutôt faible (η^2_p = .145).

Tableau 22 –Analyse de la variance des indicateurs GSR en fonction du type d'interface et du style de présentation

Sources et mesures	dll	F	η^2_p
Conductivité de la peau moyenne			
Type d'interface (I)	2	0,128	-
Style de présentation (P)	1	0,715	-
I x P	2	1,242	-
Erreur	42		
Conductivité de la peau relative			
Type d'interface (I)	2	**3,553***	.145
Style de présentation (P)	1	0,404	-
I x P	2	1,123	-
Erreur	42		

*p > .05

La figure 110 illustre l'augmentation de la conductivité de peau relative pour les conditions « *Chat 1/3* » et « *Chat 2/3* ». Néanmoins, il est difficile d'interpréter ces résultats en termes

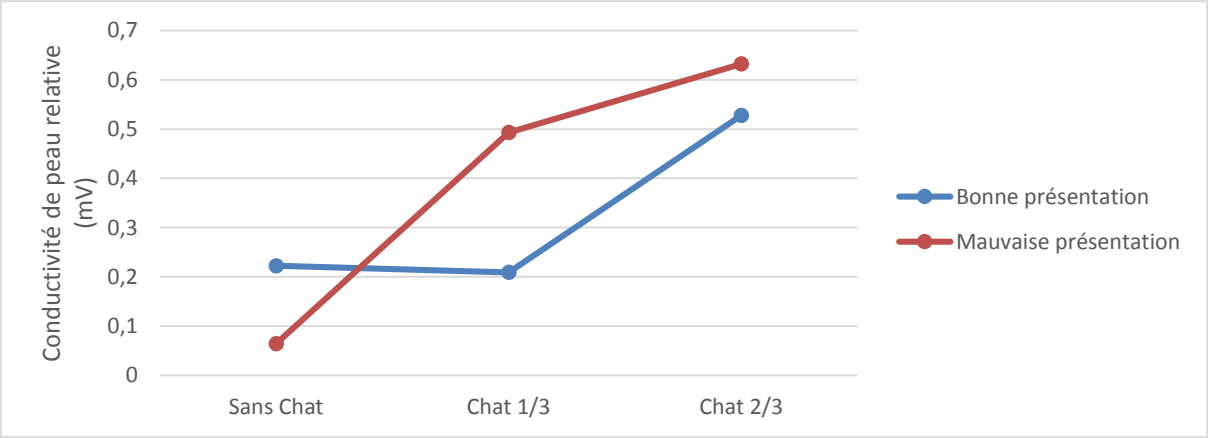


Figure 110 – Conductivité de la peau relative en fonction du type d'interface et du style de présentation

d'augmentation de l'immersion car les conditions avec chat sont celles où les utilisateurs ont eu l'opportunité d'utiliser le clavier, pouvant ainsi parasiter la capture de la conductance de la peau. De même il est possible que les corrélations constatées entre les mesures de la conductance de la peau et le questionnaire d'immersion soient dues à une variable tiers. En effet, la frappe au clavier a peut-être été impactée négativement par un niveau élevé d'immersion concentrée, ce qui peut entrainer par suite une baisse de la conductance de la peau. Ainsi, il est difficile de trancher entre une explication physique (ex : le mouvement des mains) et une explication cognitive (ex : un engagement cognitif). De plus, les données actuelles ne nous permettent pas de trancher entre ces deux hypothèses. En effet, si une différence significative existe en terme de conductivité de la peau relative entre les conditions avec et sans Chat ($t(88) = 2{,}373$, $p = .02$), on constate que le nombre de commentaires au clavier à un impact faible et non significatif sur la conductivité de la peau relative ($r = .191$ $p = .14$).

Enfin, même si la conductance de la peau a été normalisée entre les participants en prenant en compte le niveau de base individuel, l'amplitude de variation naturelle entre les participants ne l'a pas été. La Figure 111 montre l'enregistrement de deux utilisateurs ayant assisté à la même condition expérimentale et dont les amplitudes des réponses électrodermales sont complétement différentes l'une de l'autre. Sur les données des 45 participants, on note une amplitude maximum moyenne de l'ordre de 4,77mV ($\sigma = 3,87$ mV). De grandes disparités ainsi sont constatées, avec une variation minimale de 0,68mV (S23) et une variation maximale de 16,92mV, soit 25 fois plus. Cette grande disparité interindividuelle en terme d'amplitude a certainement nuit à la qualité de la mesure en brouillant le pouvoir discriminant de la conductivité de la peau pour évaluer l'immersion.

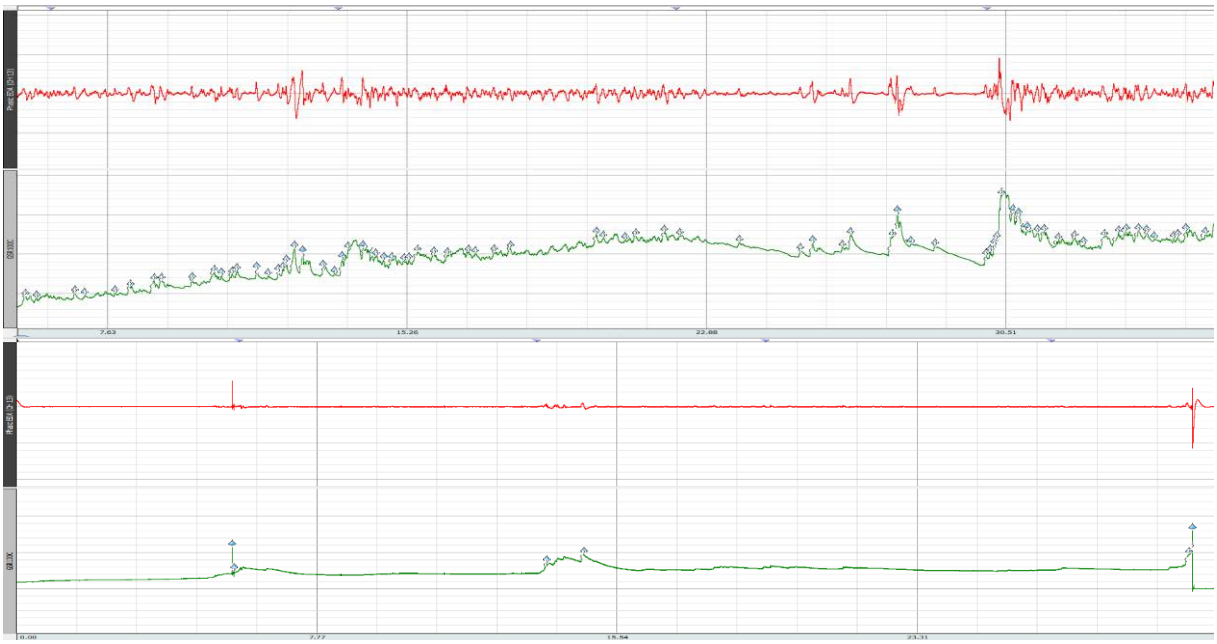


Figure 111 – Réponses électrodermales de deux participants

Expressions verbales et corporelles

Les analyses statistiques sur les données vidéo ont été réalisées sur 15 utilisateurs et pour les deux présentations. Pour cela, 5 juges ont évalué les données pour connaitre le potentiel des expressions faciales et posturales à nous renseigner sur le niveau d'immersion des utilisateurs. Deux dimensions du modèle de l'absorption cognitive ont été testées : la dimension du plaisir accru (plaisir) et l'immersion concentrée (immersion). Un débriefe qualitatif des experts sur leurs méthodologies est disponible en annexe 3D.

La fiabilité inter-juges, mesurée par le coefficient de corrélation interclasse (ou CCI) est bonne pour le jugement du plaisir (CCI = .766) et correcte pour le jugement de l'immersion (CCI = .518). Il en est de même pour l'indice de confiance auto-déclaré du jugement du plaisir (CCI = .644) et de l'immersion (CCI = 601). On constate donc que le jugement obtenu à partir des évaluations des juges est fiable, et les indices de confiance donnés pour chacun des jugements sont assez similaires d'un juge à l'autre. On ne constate pas de corrélation significative entre les jugements vidéo du plaisir et de l'immersion (r = .334, n.s.), même après une pondération des scores par les indices de confiance (r = .300, n.s.).

Pour le jugement vidéo du plaisir, on ne constate pas de corrélation significative avec la dimension de plaisir accru et le score d'immersion total, avec ou sans pondération (Table 23). Par contre, pour le jugement vidéo de l'immersion, on constate une corrélation modérée avec la dimension d'immersion concentrée et une corrélation importante avec le score d'immersion total, que ce soit avec ou sans pondération (Table 24).

Tableau 23 – Corrélations entre le jugement vidéo du plaisir et les échelles du questionnaire d'immersion

	Jugement vidéo du Plaisir	Jugement vidéo du Plaisir (pondéré par l'indice de confiance)
Plaisir accru	.221 (n.s.)	.208 (n.s.)
Score d'immersion total	.181 (n.s.)	.163 (n.s.)

Tableau 24 – Corrélations entre le jugement vidéo de l'immersion et les échelles du questionnaire d'immersion

	Jugement vidéo de l'immersion	Jugement vidéo de l'immersion (pondéré par l'indice de confiance)
Immersion concentrée	**.353***	**.351***
Score d'immersion total	**.540****	**.525****

*p < .05 **p<.001

Une analyse de la variance (Tableau 25), nous montre que seul le jugement vidéo de l'immersion sans pondération a été capable de discriminer significativement entre les styles de présentation ($F(1,42) = 3,553$, $p = .047$, $\eta^2_p = .254$).

Tableau 25 –Analyse de la variance des jugements vidéo en fonction du style de présentation

Mesures	**dll**	**F**	η^2_p
Jugement vidéo du Plaisir			
Non pondéré	1,14	2,284	-
Pondéré par l'indice de confiance	1,14	1,655	-
Jugement vidéo de l'immersion			-
Non pondéré	1,14	**4,760***	.254
Pondéré par l'indice de confiance	1,14	4,478	-

*p > .05

Traces écrites

Les analyses statistiques sur les traces écrites ont été réalisées sur 30 utilisateurs et pour les deux présentations. Pour cela, 5 juges ont analysé les données pour connaitre le potentiel des traces écrites pour nous renseigner sur le niveau d'immersion des utilisateurs. Comme pour l'analyse des vidéos, deux dimensions du modèle de l'absorption cognitive ont été testées : la dimension du plaisir accru (plaisir) et l'immersion concentrée (immersion).

La fiabilité inter-juges, est correcte pour le jugement du plaisir (CCI = .640) et très bonne pour le jugement de l'immersion (CCI = .824). Il en est de même pour l'indice de confiance auto déclaré du jugement du plaisir (CCI = .853) et de l'immersion (CCI = 851). On constate donc que le jugement obtenue à partir des évaluations des juges est fiable, et tout particulièrement le jugement de l'immersion. Les indices de confiance donnés pour chacun des jugements sont très similaires d'un juge à l'autre, ce qui signifie que les juges ont retrouvé les mêmes difficultés/facilités lors de l'évaluation des verbalisations écrites. De plus, on constate une

corrélation significative entre les jugements du plaisir et de l'immersion, avec ou pondération par l'indice de confiance

Pour le jugement du plaisir et de l'immersion, on observe une série de corrélations significatives avec les dimensions respectives du questionnaire d'immersion, avec ou sans pondération (Table 23 et 24).

Tableau 26 – Corrélations entre le jugement des traces écrites du plaisir et les échelles du questionnaire d'immersion

	Jugement des traces écrites du Plaisir	Jugement des traces écrites du Plaisir (pondéré par l'indice de confiance)
Plaisir accru	**.421****	**.407****
Score d'immersion total	**.382***	**.360***

*p < .01 **p<.001

Tableau 27 – Corrélations entre le jugement des traces écrites de l'immersion et les échelles du questionnaire

	Jugement des traces écrites de l'immersion	Jugement des traces écrites de l'immersion (pondéré par l'indice de confiance)
Immersion concentrée	**.462****	**.434****
Score d'immersion total	**.381***	**.351***

*p < .01 **p <.001

Néanmoins, une analyse de la variance (Tableau 25), nous montre que seuls les jugements des traces écrites du plaisir, avec et sans pondération, ont été capables de discriminer significativement entre les styles de présentation ($F(1,28) = 14,647$, $p = .001$, $\eta^2_p = .343$; $F(1,42) = 13,122$, $p = .001$, $\eta^2_p = .319$).

Tableau 28 – Analyse de la variance des jugements des traces écrites en fonction du type d'interface et du style de présentation

Sources et mesures	**dll**	**F**	η^2_p
Jugement des traces écrites du plaisir			
Type d'interface (I)	1	2,688	.319
Style de présentation (P)	1	**13,122***	-
I x P	1	0,078	-
Erreur	28		
Jugement des traces écrites du plaisir (pondéré)			
Type d'interface (I)	1	1,905	-
Style de présentation (P)	1	**14,647***	.343
I x P	1	0,004	-
Erreur	28		
Jugement des traces écrites de l'immersion			
Type d'interface (I)	1	3,028	-
Style de présentation (P)	1	0,838	-
I x P	1	0,346	-
Erreur	28		
Jugement des traces écrites de l'immersion (pondéré)			
Type d'interface (I)	1	1,943	-
Style de présentation (P)	1	0,712	-
I x P	1	0,168	-
Erreur	28		

*p > .001

Analyse mono-mesure (Expérience 2)

Méthodologie

Plan expérimental

Pour tester la fiabilité des mesures, un plan expérimental à deux facteurs (2 x 3), manipulant le « *Type d'interface* » (« *Bonne interface* », « *Moyenne interface* », « *Mauvaise interface* ») et le « *Style de présentation* » (« *Bonne présentation* » et « *Mauvaise présentation* ») a été élaboré (Figure 111). Chaque participant a été assigné aléatoirement à une des trois interfaces (condition emboitée) et à chacun des deux styles de présentation, successivement (condition croisée). Pour contrôler l'effet d'ordre, le style de présentation a été contrebalancé entre les participants.

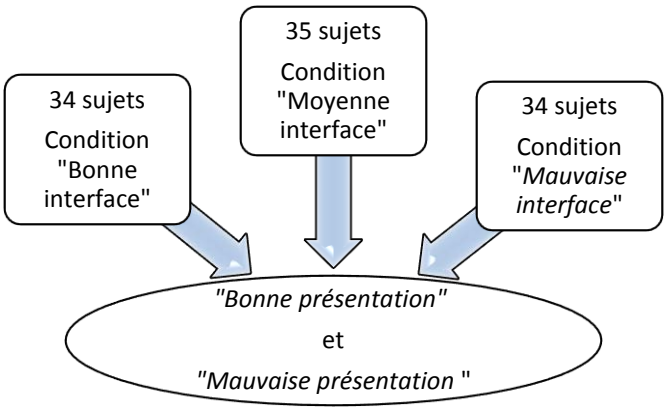


Figure 111 – Plan expérimental et répartition des utilisateurs

Artefact de test

Composition des interfaces en fonction des antécédents de l'immersion

Les nouvelles interfaces ont été conçues en se basant sur trois facteurs influençant l'immersion : le confort visuel, la distraction et le contrôle de l'information. Le confort visuel peut être manipulé par la taille (Neuman, 1990; Reeves, Lombard, & Melwani, 1992) et la qualité de l'image (Bocker & Muhlbach, 1993) en augmentant le réalisme et la capacité à s'engager émotionnellement et cognitivement. Les détracteurs, d'ordres visuels ou sonores, sont des antécédents défavorisant l'immersion car ils perturbent la concentration et donc l'attention (Jenett, 2010; Walker & Lindsay, 2006). Enfin, un environnement peut simuler la curiosité de l'utilisateur s'il fournit un niveau optimal d'informations : elles doivent être ni trop compliquées ni trop simples par rapport aux connaissances existantes de l'utilisateur. Le contrôle de l'information permet donc de faciliter l'implication de l'individu et favorise sont immersion cognitive (Pace, 2003). En se basant sur ces facteurs, trois interfaces avec des niveaux d'immersion différents ont été conçues :

- Une interface ennuyeuse et frustrante intégrant des antécédents qui diminuent l'immersion
- Une interface immersive comprenant les antécédents reconnus comme améliorant l'immersion
- Une interface intermédiaire, dite moyennement immersive, qui est un compromis entre les deux interfaces précédentes

Le tableau 29 résume la composition des trois interfaces par rapport aux antécédents de l'immersion dégagés par l'état de l'art.

Tableau 29 – Composition des trois interfaces par rapport aux antécédents de l'immersion dégagés par l'état de l'art

		Mauvaise interface	Moyenne interface	Bonne interface
Antécédents de l'immersion	**Confort visuel**	Petite taille de fenêtre	Moyenne taille de fenêtre	Grande taille de fenêtre
		Qualité et débit d'image de la présentation faible	Qualité et débit d'image de la présentation normal	Bonne qualité d'image de la présentation et haut débit
	Distracteurs	Arrivé des images des participants en pop-up avec bruit sonore de connexion	Arrivé des images des participants en pop-up avec bruit sonore de connexion	Absence de pop-up et de bruit de connexion
		Bruits sonores pour chaque intervention des participants et des nouveaux événements	Notification par clignotement des onglets pour marquer la présence de nouveaux événements	Notification graphique et discrète pour marquer la présence de nouveaux événements
		Pop-up aléatoire de messages d'erreur : « *Cliquez sur Ok pour continuer* »	Absence de messages d'erreur	Absence de messages d'erreur
	Contrôle de l'information	Absence de l'onglet « Informations supplémentaires »	Présence de l'onglet « Informations supplémentaires »	Présence de l'onglet « Informations supplémentaires »
		Module de Chat imposé	Affichage alternatif du Chat ou des informations supplémentaires	Affichage/masquage du Chat ou des informations supplémentaires

Maquettes interactives

L'ancienne version de la maquette a été améliorée en intégrant différentes fonctionnalités permettant de mieux manipuler l'immersion. D'autres fonctionnalités, comme la saisie dans le module de Chat, ont été supprimées, car elles n'étaient plus utiles (pas de recueil des traces écrites dans l'expérience 2) et étaient une source probable d'interférences pour la capture des signaux physiologiques.

La nouvelle maquette comprend quatre parties :

- Une fenêtre principale de diffusion de flux vidéo pour l'affichage de la vidéoconférence
- Un onglet « audience » avec deux fenêtres : une fenêtre de chat, simulant une interaction entre les participants (pour faire ressentir à l'utilisateur la présence sociale des autres auditeurs à distance) ; une fenêtre listant les différents auditeurs connectés à distance (permettant de connaître leur prénom et l'entreprise d'appartenance)
- Un onglet « informations supplémentaires » ayant pour rôle d'introduire la séance et de compléter certaine notion propre à la conférence par des schémas ou des définitions complexes. Ces informations sont consultables selon l'intérêt du l'auditeur.

Les figures 112, 113 et 114 représentent les trois différentes interfaces de la maquette et quelques-unes de leurs particularités (représentées par les panneaux).

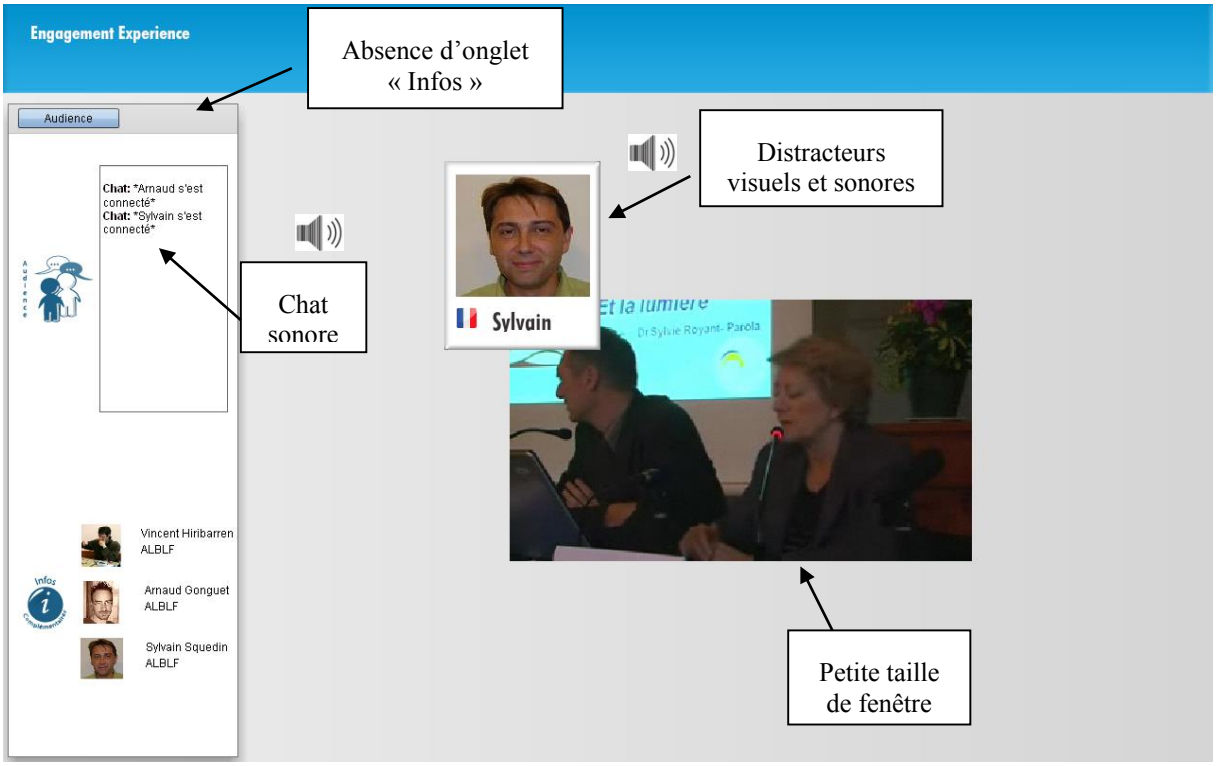


Figure 112 – Capture d'écran de l'artefact de test dans la condition « Mauvaise interface »

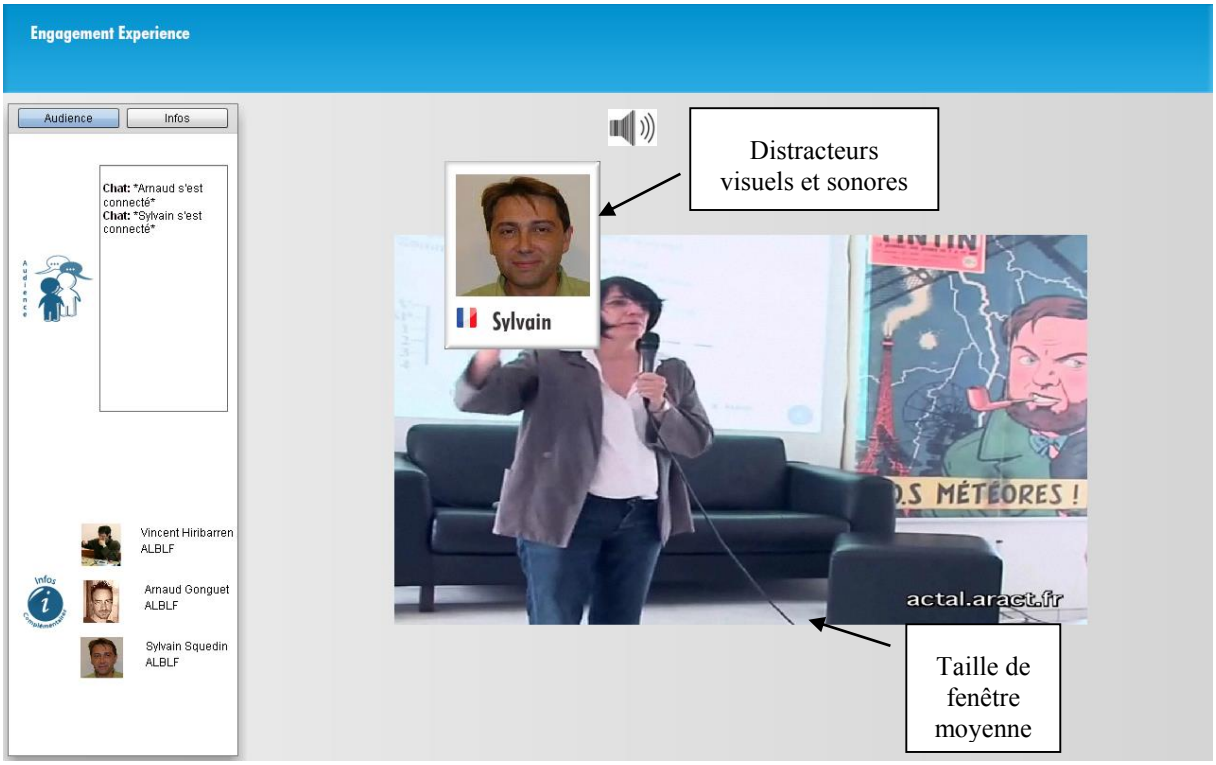


Figure 113 – Capture d'écran de l'artefact de test dans la condition « Moyenne interface »

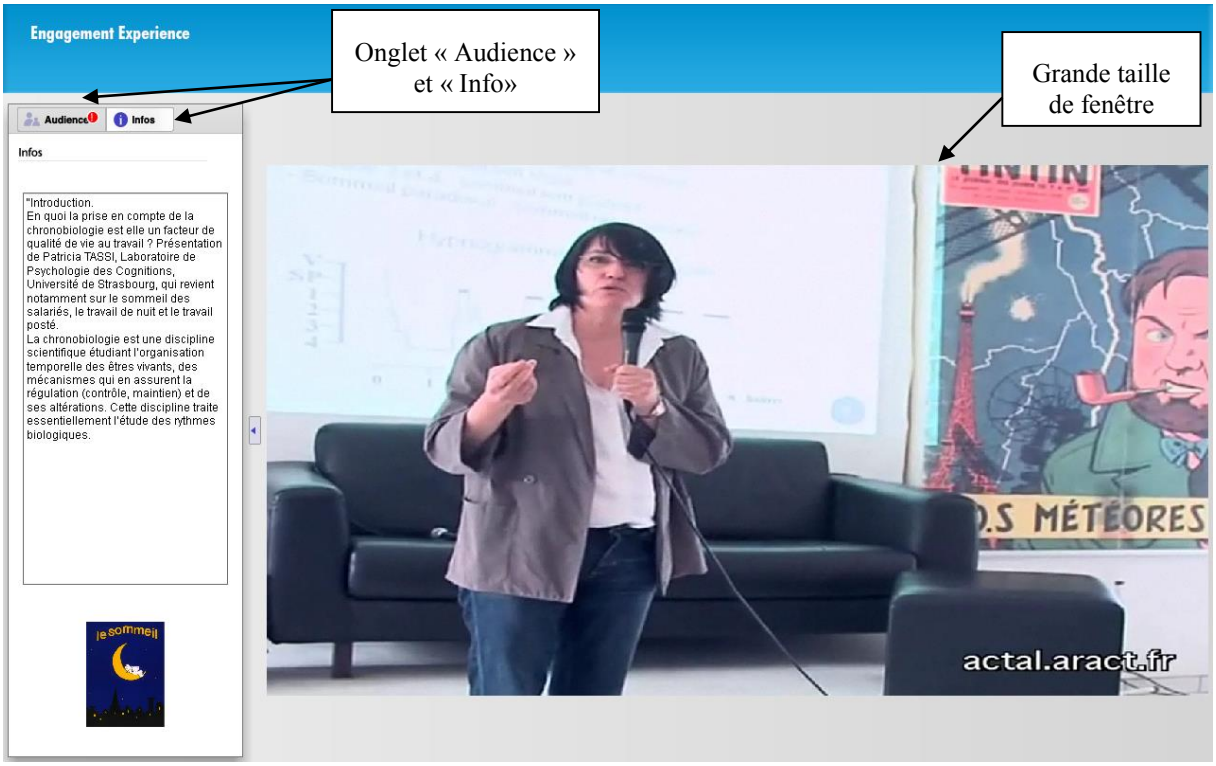


Figure 114 – Capture d'écran de l'artefact de test dans la condition « Bonne interface »

Présentations vidéos

Afin de manipuler le niveau d'immersion des participants, deux vidéoconférences de même durée ont été choisies. Elles se différencient par leurs qualités didactiques et pédagogiques, ainsi que par leurs qualités d'édition. Les deux présentations portent sur le sommeil. La vidéo « *Bonne présentation* » (7'19'') est une introduction aux différents stades du sommeil. L'orateur est mobile et dynamique. La présentation intègre des diapositives explicatives rendant la conférence accessible et intéressante. Le sujet n'est ni trop facile ni trop difficile et constitue donc un défi motivant à l'écoute. La tâche est en soi source de satisfaction et elle n'est donc pas perçue comme une corvée (Figure 115). La vidéo « Mauvaise présentation » (6'39'') a pour thème la luminothérapie et le sommeil. La présentation est confuse, compliquée et ennuyeuse. L'orateur est statique et monotone. Il n'y a pas de présentation de diapositive. La tâche d'écoute constitue un défi démotivant et peut être perçue comme une corvée (Figure 116).

Figure 115 – Vidéo « Bonne présentation »

Figure 116 – Vidéo « Mauvaise présentation »

Matériel physique et configuration logicielle

Figure 117 − Système d'acquisition BIOPAC MP150 avec amplificateur GSR100C et PPG100C, et électrodes EL507»

Figure 118 – Earclip TSD200A et Eyetracker Tobii T60

Sur le plan technique, deux ordinateurs, un Eyetracker Tobii et un système d'acquisition physiologique BIOPAC ont été mobilisés pour l'expérience (Figure 117 et 118). Un PC *Sony VAIO* a été utilisé comme ordinateur de test : il est connecté à l'écran Tobii, héberge l'enregistrement de la session via *Tobii Studio* et exécute la maquette interactive. Un PC *Asus ROG G73Sw* a été utilisé comme poste d'acquisition et de contrôle du signal *GSR et PPG* via le logiciel *Acqknowledge*. L'ensemble des postes informatiques et le *BIOPAC MP150* ont été reliés par un *Switch Ethernet 3Com GSU05* via des câbles *RJ45*. Enfin, un thermomètre a permis de contrôler la température de la salle de test avant chaque passation.

Mesures

Auto-rapporté (Questionnaire d'immersion)

Le questionnaire immersion en 9 item de l'expérience 1 a été réutilisé pour servir de mesure de comparaison. Deux autres mesures auto-rapportées ont toutefois été ajoutées pour approfondir l'analyse de ce type de mesure : une mesure intuitive de la durée de la vidéoconférence et une estimation du niveau d'immersion en 1 item. Ces données ont été recupérées pour les 103 participants.

La mesure intuitive du temps écoulé se base sur toute une série d'expériences qui montre que l'on a tendance à sous-estimer la durée d'une expérience engageante (émotionnellement et cognitivement) alors que, au contraire, l'on surestime en moyenne la durée une expérience ennuyante et peu engageante (Hancock & Weaver, 2005; Hertzum & Holmegaard, 2013; Loftus, Schooler, Boone, & Kline, 1987; Rau, Peng, & Yang, 2006; Sanders & Cairns, 2010;

Wood, Griffiths, & Parke, 2007). Cet indicateur permet un contrôle supplémentaire sur la mesure de la dissociation temporelle. Cette mesure s'est appuyé sur la question : « *Combien de temps a, selon vous, duré la vidéoconférence ?* » et une échelle en 5 points, allant de « *5 minutes* » à « *9 min* ». Elle a ensuite été corrigée pour traduire l'écart en seconde avec la durée réelle de la vidéoconférence en question.

Une mesure mono-item de l'immersion, s'inscrivant dans la tendance des questionnaires réduits (Schweizer, 2011), a été utilisée pour tester un moyen peu coûteux et intrusif de suivre l'expérience des utilisateurs par une évaluation rapide et autonome de ces derniers (Lottridge, 2009). Cette mesure s'est appuyée sur la question « *Sur une échelle de 1 à 10, à quel niveau étiez-vous immergé dans la vidéoconférence ?* » et une échelle ordinale en dix points.

Conductance de la peau (GSR)

La capture de la conductance de la peau a été réalisée sur le même modèle que l'expérience 1, c'est à dire grâce au système d'acquisition de données BIOPAC MP150, d'un amplificateur du signal galvanique GSR100C et de deux Electrodes EL507. Le protocole général a toutefois été repensé pour que le participant n'ait plus à utiliser le clavier, et ainsi limiter les perturbations physiques pouvant nuire à la qualité du signal physiologique recueilli.

Cinq indicateurs ont été recueillis. Il s'agit de la conductivité de la peau moyenne (GSR_{MOY}), la Baseline (GSR_{BASE}), la conductivité de la peau relative ($GSR_{REL.}$) et deux indicateurs supplémentaires : la conductivité de la peau maximum (GSR_{MAX}) et la GSR normalisée (GSR_{NORM}). La conductivité de la peau maximum a été obtenue à la fin de la passation, en faisant jouer le participant à un rapide jeu de vitesse et de réflexion, dans le but d'obtenir l'amplitude maximale de la GSR. La GSR normalisée a ensuite été calculée par la formule de normalisation suivante (M. E. Dawson et al., 2007) :

$$GSR_{NORM} = \frac{GSR_{MOY} - GSR_{BASE}}{GSR_{MOY} - GSR_{MAX}}$$

Cet indicateur a été calculé pour contrôler la différence d'amplitude naturelle de conductivité dermale propre à chaque participant.

Les données GSR ont été capturées pour 96 utilisateurs, soit 31 pour la condition « Bonne interface », 32 pour la condition « Moyenne interface » et 33 pour la condition « Mauvaise interface ». Les données pour 7 utilisateurs ont été rejetées, soit parce que l'enregistrement était plat (mauvais contact probable des électrodes) ou soit parce le signal était fortement parasité (mouvement importants).

Rythme cardiaque (PPG)

La fréquence cardiaque a été obtenue par la mesure de la pulsation sanguine (BVP). Pour cela, un système d'acquisition de données BIOPAC MP150, un amplificateur du signal PPG100C et un Earclip TSD200A a été utilisé. Cette mesure s'est basée sur la technique de la photopléthysmographie (PPG), qui permet de mesurer la fréquence cardiaque par un photo-capteur de détection du volume sanguin périphérique. En utilisant la réflexion cutanée de la lumière infrarouge au niveau d'une région fine, tel que le lobe de l'oreille, il est possible inférer la fréquence cardiaque en mesurant la quantité de sang sur des intervalles de temps précises. 3

indicateurs ont ainsi été recueillis : la moyenne du rythme cardiaque (PPG_{MOY}), la Baseline (PPG_{BASE}), et le rythme cardiaque relatif (PPG_{REL}).

La moyenne du rythme cardiaque (en battement par minute) a été calculée par la moyenne des données physiologiques obtenue durant toute la période de la vidéoconférence. Ces données ont été filtrées et corrigées. En effet, le signal brut a été parasité par les nombreux mouvements de tête. Tous les artefacts de mesures ont été retirés manuellement sur les courbes de données. Ces données ne prennent pas en comptes la variabilité interindividuelle du rythme cardiaque. La Baseline a été calculée par la moyenne des données physiologiques obtenue durant la période de relaxation. Elles ont permis d'avoir le rythme cardiaque des participants avant l'expérience de vidéoconférence. Enfin, le rythme cardiaque relatif a été calculé par la différence entre la moyenne du rythme cardiaque et la Baseline (PPG_{MOY} - PPG_{BASE}). Ces données prennent en comptes la variabilité interindividuelle du rythme cardiaque.

Les données PPG ont été capturées pour 66 utilisateurs, soit 22 pour la condition « Bonne interface », 23 pour la condition « Moyenne interface » et 21 pour la condition « Mauvaise interface ». Les données pour 37 utilisateurs ont été rejetées parce que l'enregistrement était trop parasité. Le critère de rejet était de 20 HR_Reset, ce qui correspond au nombre de décrochage de l'algorithme de détection du rythme cardiaque à partir des données brutes. En effet, même après une suppression manuelle des artefacts sur les courbes physiologiques, ces enregistrements étaient trop bruités pour être utilisés sans risque dans l'étude.

Mouvements oculaires

Les mouvements oculaires des sujets ont été mesurés par un eye-tracker Tobii T60. Cet écran a capturé les images des yeux à partir d'une caméra infrarouge pour déterminer la direction du regard sur l'écran. Il a été calibré avant chaque début de séance par les suivis oculaires de l'utilisateur d'un point rouge sur l'écran.

Deux indicateurs simples ont été recueillis : le nombre de fixations et la durée totale de fixation. A partir de ces deux indicateurs, la durée moyenne par fixation a été calculée. Les deux conférences ayant des durées légèrement différentes (7'19'' et 6'39''), les données pour les deux vidéoconférences ont été normalisées pour correspondre toutes les deux à une durée de 7 minutes. Deux indicateurs d'évolution du comportement oculaire ont également été calculés à partir de ces indicateurs corrigés : l'évolution du nombre de fixations et l'évolution de la durée moyenne par fixation. Pour calculer ces indicateurs, l'enregistrement oculaire concernant une vidéoconférence a été divisé en 7 segments d'une minute, puis, a été utilisé pour le calcul de contraste linéaire. Ce contraste nous permet d'avoir la tendance linaire au cours du temps, c'est-à-dire, une augmentation, une baisse ou une stagnation au cours du temps. Les coefficients d'un contraste linéaire avec 7 segments sont les suivants : *-3, -2, -1, 0, 1, 2* et *3*. Ainsi, le contraste linéaire du nombre de fixations pour le participant 1 durant la « *bonne présentation* » est de :

-3(184,84)-2*(104,52)-1*(146,13)+0*(140,32)+1(140,32)+2*(109,35)+3*(103,55) = -240,02* soit une diminution du nombre de fixations au cours du temps. Un contraste négatif indique une baisse, un nombre positif une augmentation, et un nombre proche de 0 une stagnation de la tendance linéaire.

Les données oculaires ont été capturées pour 94 utilisateurs, soit 31 pour la condition « Bonne interface », 32 pour la condition « Moyenne interface » et 31 pour la condition « Mauvaise interface ». Les données pour 9 utilisateurs ont été rejetées. Le critère de rejet était une qualité d'enregistrement (calculée par Tobii studio) inférieure à 66% ou des données oculaires manquantes durant plus d'une minute.

Recrutement et procédure

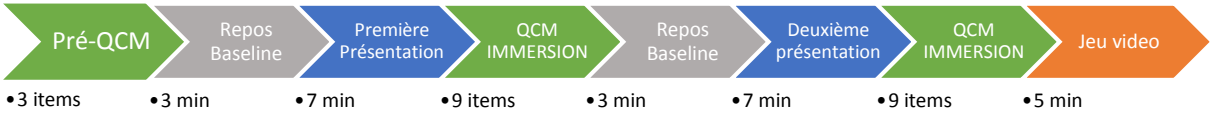


Figure 119 – Ordre de réalisation du test

Les participants ont été recrutés au sein des Bell Labs. Leur langue courante a été contrôlée préalablement au recrutement, car elle est susceptible d'influer sur la compréhension. Par précaution, seules des personnes parlant couramment le français depuis plus de dix ans ont été recrutées. De plus, pour ne pas biaiser les enregistrements de la conductance de la peau, les participants ne doivent pas avoir les mains excessivement fripées, excessivement froides (doigts virant au bleu), et ils ne doivent pas présenter de troubles de la sudation tels que l'hypersudation ou encore la maladie de Raynaud (doigts virant au blanc). De ce fait, ces prérequis ont été systématiquement contrôlés par une observation informelle des mains des sujets lors des contacts de recrutement.

Avant chaque passation, la température ambiante a été systématiquement contrôlée pour vérifier si celle-ci était comprise entre 22°C et 25°C. Au début de la passation, le participant a été assigné aléatoirement à l'une des trois conditions basées sur l'interface : « Bonne interface », « Moyenne interface » et « Mauvaise interface ». Les participants ont été ensuite installés, les capteurs GSR ont été branchés et la vérification de l'assise du sujet et de la posture correcte de la main (paume orientée vers le haut) ont été réalisées. Puis, le participant a été accompagné pour la lecture des consignes générales et le casque audio a été mis en place (en faisant attention au capteur PPG). Les participants ont été informés qu'ils allaient devoir regarder deux vidéoconférences. Puis, le logiciel de capture physiologique Acqknowledge a été lancé pour vérifier le calibrage en parallèle de celui de l'eye-tracker. Ensuite, le participant a bénéficié de cinq minutes de repos pour établir la Baseline de la GSR. Puis, la première vidéo a été diffusée. A la fin de la première vidéo, le questionnaire d'immersion a été présenté, suivi par cinq nouvelles minutes de repos pour rétablir la Baseline de la GSR. Enfin, la deuxième vidéo a été diffusée, suivit par la deuxième présentation du questionnaire d'immersion. Avant de finir le test, on demande au participant de jouer à un jeu de vitesse et de réflexion dans le but d'obtenir l'amplitude maximale de la GSR. On remercie le sujet pour sa participation et on lui propose, si cela l'intéresse, de lui présenter les courbes physiologiques obtenues. On raccompagne le sujet vers la sortie. Le protocole opératoire et les consignes complètes sont disponibles en annexe 3C. La figure 120 présente la configuration de la salle de test lors d'une passation.

Figure 120 – Configuration de la salle de test

Résultats

Descriptif de l'échantillon et des conditions de passation

103 sujets, 18 femmes et 42 hommes ont participé à l'expérimentation, soit 34 pour la condition « Bonne interface », 35 pour la condition « Moyenne interface » et 34 pour la condition « Mauvaise interface ». L'âge moyen des participants était de 32 ans ($\sigma = 9{,}22$). Plus de la moitié des sujets (58%) avaient un niveau d'études équivalent à BAC+5, et plus d'un cinquième (21%) des sujets un niveau d'études équivalent à BAC+8. La durée moyenne des passations était de 35 minutes par sujet et la température ambiante de la salle de passation du test était en moyenne de 23,4°C, ($\sigma = 0{,}74$).

Mesure auto-rapportée (Questionnaire d'immersion)

Les analyses statistiques sur le questionnaire d'immersion ont été réalisées sur les 103 participants à l'expérience. La consistance interne du questionnaire, mesurée par le coefficient alpha de Cronbach, est bonne ($\alpha = .868$), ainsi que pour les sous-échelles de plaisir accru ($\alpha = .881$) et d'immersion concentrée ($\alpha = .847$). La consistance interne de la sous-échelle de dissociation temporelle ($\alpha = .571$) est passable. Les corrélations entre les différentes sous-échelles du questionnaire nous montrent des liens forts et très significatifs ($p > .001$) entre les trois dimensions mesurées (tableau 30).

Tableau 30 – Corrélations entre les trois sous-échelles du questionnaire d'immersion

	Plaisir accru	Immersion concentrée
Immersion concentrée	**.484****	-
Dissociation temporelle	**.577****	**.401****

** p < .01

Une analyse de la variance (Tableau 31), nous montre que le questionnaire d'immersion a permis de discriminer significativement entre la « *bonne* » et la « *mauvaise* » présentation ($F(1,100) = 105{,}203$, $p > .000$, $\eta^2_p = .513$) et entre les types d'interfaces ($F(2,100) = 6{,}798$, $p = .002$, $\eta^2_p = .120$). Pas d'effet d'interaction entre le style de présentation et le type d'interface n'a été détecté ($p < .05$).

Tableau 31 – Analyse de la variance du questionnaire d'immersion en fonction du type d'interface et du style de présentation

Sources	**dll**	**F**	η^2_p
Type d'interface (I)	2	**6,798***	.120
Style de présentation (P)	1	**105,203****	.513
I x P	2	0,996	-
Erreur	100		

*p > .01 **p > .001

La figure 121 illustre comment le questionnaire a discriminé les interfaces et les présentations en fonction du niveau d'immersion déclaré des participants.

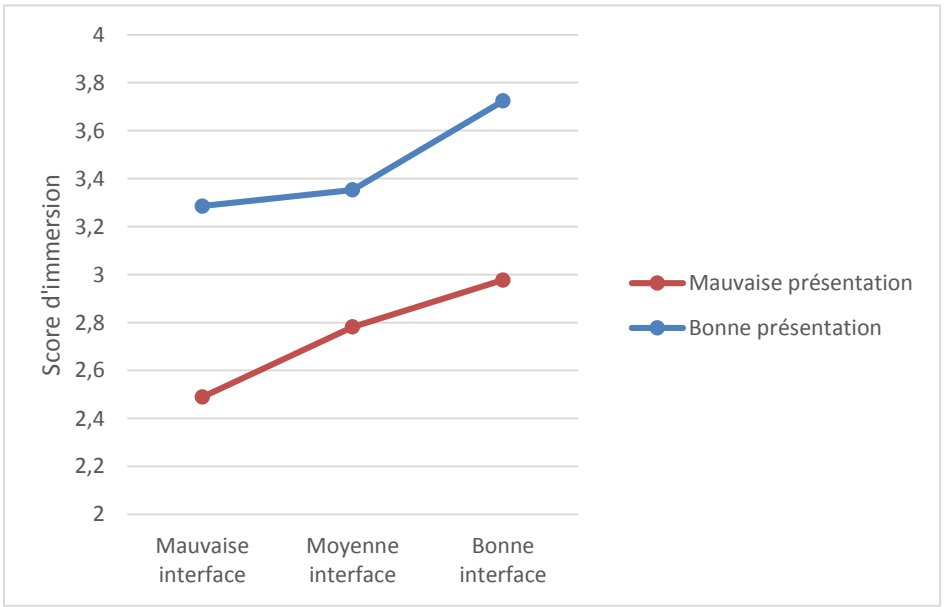


Figure 121 – Score d'immersion en fonction du type d'interface et du style de présentation

Pour l'estimation subjective de la durée de la conférence, on observe des corrélations faibles à modérés avec le questionnaire d'immersion et ses sous-échelles (table 32). Néanmoins, cet indicateur n'a pas permis de discriminer significativement entre les types d'interfaces et les styles de présentation (Table 33).

Tableau 32 – Corrélations entre l'estimation subjective de la durée de la conférence et les échelles du questionnaire d'immersion

	Plaisir accru	Immersion concentrée	Dissociation temporelle	Score d'Immersion total
Estimation subjective de la durée de la conférence	**-.335****	**-.248***	**-.347****	**-.381****

*p > .01 **p > .001

Tableau 33 – Analyse de la variance de l'estimation subjective de la durée de la conférence en fonction du type d'interface et du style de présentation

Sources	**dll**	**F**	η^2_p
Type d'interface (I)	2	0,996	-
Style de présentation (P)	1	3,345	-
I x P	2	2,702	-
Erreur	100		

Pour l'estimation la mesure mono-item de l'immersion, on observe une corrélation très forte avec le score total du questionnaire d'immersion (*r* = .786), ainsi que des corrélations fortes avec toutes ses sous-échelles (table 34).

Tableau 34 – Corrélations entre la mesure de l'immersion mono-item et les échelles du questionnaire d'immersion

	Plaisir accru	Immersion concentrée	Dissociation temporelle	Score d'Immersion total
Mesure de l'immersion mono-item	**.731****	**.672****	**.495****	**.786****

**p > .001

Une analyse de la variance (Tableau 35), nous montre que la mesure d'immersion mono-item a permis de discriminer significativement entre la « *bonne* » et la « *mauvaise* » présentation ($F(1,100) = 153,942$, $p > .000$, $\eta^2_p = .606$) et entre les types d'interfaces ($F(2,100) = 13,048$, $p > .000$, $\eta^2_p = .207$), soit une sensibilité de discrimination supérieure à celle observée pour le questionnaire. Pas d'effet d'interaction entre le style de présentation et le type d'interface n'a été détecté ($p < .05$).

Tableau 35 – Analyse de la variance du questionnaire d'immersion en fonction du type d'interface et du style de présentation

Sources	**dll**	**F**	η^2_p
Type d'interface (I)	2	**13,048****	.207
Style de présentation (P)	1	**153,942****	.606
I x P	2	0,348	-
Erreur	100		

**p > .001

La figure 122 nous permet de comparer les résultats obtenus par le questionnaire d'immersion en 9 items et la mesure d'immersion mono-items pour les deux vidéos de présentation et les trois niveaux d'interface. Pour cela, les scores de l'échelle en cinq points du questionnaire en 9 items ont été normalisés pour correspondre à l'échelle en 10 points de la mesure mono-item. Nous pouvons constater que la progression entre les différentes conditions est beaucoup plus linéaire avec le questionnaire mono-items et que l'amplitude entre les scores les plus bas et les plus hauts est plus importante.

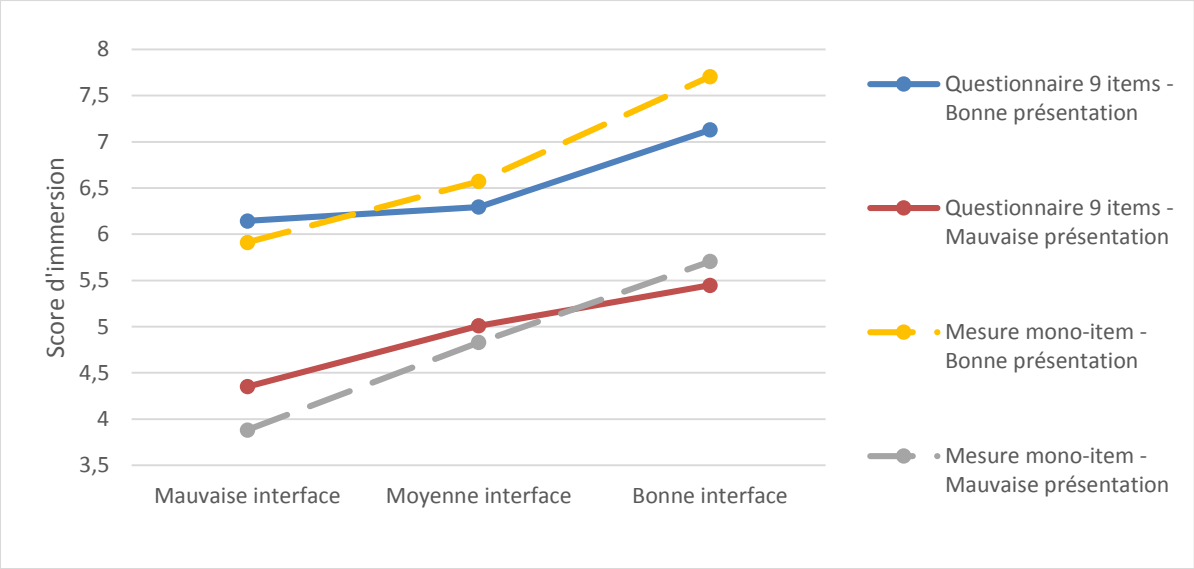


Figure 122 - Score d'immersion en fonction du type d'interface et du style de présentation pour le questionnaire d'immersion en 9 items et la mesure d'immersion mono-item

Conductance de la peau (GSR)

Les analyses statistiques sur les données physiologiques de la conductance de la peau ont été réalisées sur 66 participants. Pour la conductivité moyenne de la peau et la conductivité de la peau normalisée, aucun lien significatif avec l'échelle auto-rapportée d'immersion et ses sous-échelles n'a été trouvé. Pour la conductivité de la peau relative, on constate une corrélation négative faible avec le score d'immersion total (r = -.145, p = .046).

Tableau 36 – Corrélations entre les indicateurs GSR et les échelles du questionnaire

	Conductivité de la peau moyenne	Conductivité de la peau relative	Conductivité de la peau normalisée
Plaisir accru	-.093 (n.s.)	-.034 (n.s.)	.038 (n.s.)
Immersion concentrée	.050 (n.s.)	.054 (n.s.)	-.020 (n.s.)
Dissociation temporelle	.038 (n.s.)	-.075 (n.s.)	.036 (n.s.)
Score d'immersion total	-.049 (n.s.)	**-.145***	.036 (n.s.)

*p < .01

Une analyse de la variance (Tableau 37), nous montre que la conductivité de la peau relative a permis de discriminer significativement entre les types d'interface (F(2,93) = 6,325, p = .003, η^2_p = .145) et les présentations (F(1,93) = 6,102, p = .015, η^2_p = .062) avec des tailles d'effet plutôt faibles.

Tableau 37 – Analyse de la variance des indicateurs GSR en fonction du type d'interface et du style de présentation

Sources et mesures	dll	F	η^2_p
Conductivité de la peau moyenne			
Type d'interface (I)	2	0,474	-
Style de présentation (P)	1	0,279	-
I x P	2	0,887	-
Erreur	93		
Conductivité de la peau relative			
Type d'interface (I)	2	**6,325***	.120
Style de présentation (P)	1	**6,102***	.062
I x P	2	1,902	-
Erreur	93		
Conductivité de la peau normalisée			
Type d'interface (I)	2	2,028	-
Style de présentation (P)	1	0,887	-
I x P	2	1,406	-
Erreur	93		

*p > .05 **p > .01

La figure 123 illustre la variation de la conductivité de peau relative entre les types d'interface et les vidéos. On constate que la conductivité de peau relative est plus importante pour la bonne présentation que pour la mauvaise, mais, en même temps, elle diminue de la mauvaise interface à la bonne. Cela pose un problème majeur pour l'utilisation de la conductance de la peau en tant que mesure de l'immersion car sa variation n'est pas indexée linéairement avec un niveau associé d'immersion.

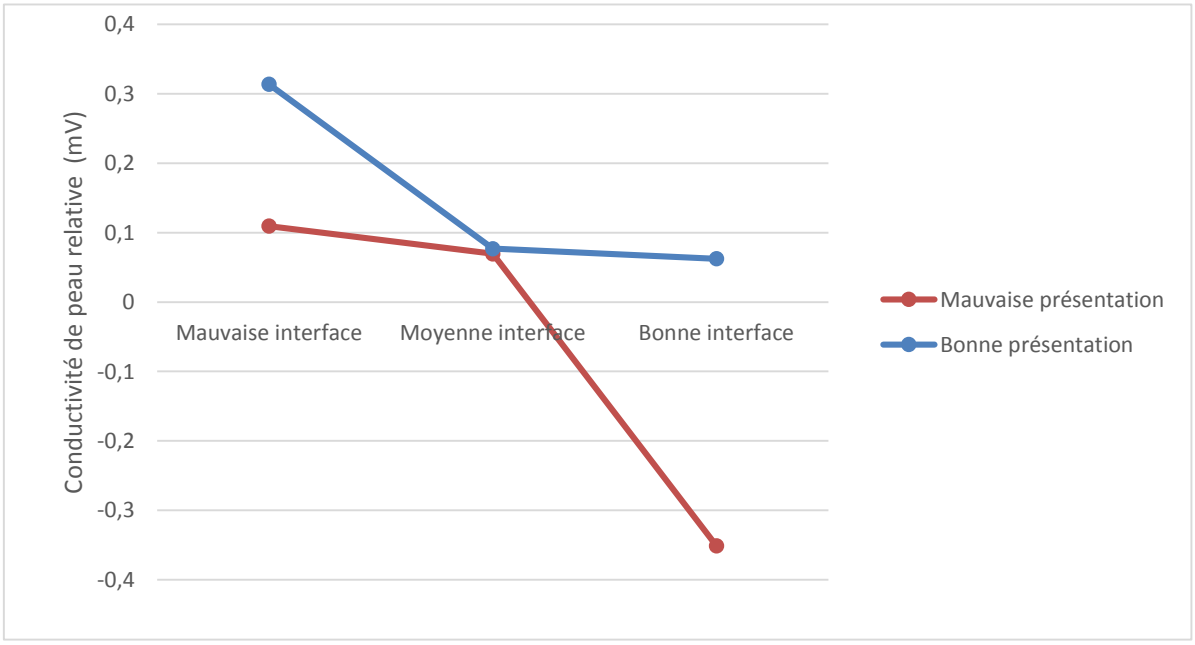


Figure 123 – Conductivité de la peau relative en fonction du type d'interface et du style de présentation

Rythme cardiaque (PPG)

Les analyses statistiques sur le rythme cardiaque ont été réalisées sur 66 participants. Pour le rythme cardiaque moyen et relatif, aucun lien significatif avec l'échelle auto-rapportée d'immersion et ses sous-échelles n'a été trouvé ($p > .05$) ; (Tableau 38).

Tableau 38 – Corrélations entre les indicateurs PPG et les échelles du questionnaire

	Rythme cardiaque moyen	Rythme cardiaque relatif
Plaisir accru	.095 (n.s.)	.124 (n.s.)
Immersion concentrée	-.007 (n.s.)	-.011 (n.s.)
Dissociation temporelle	.087 (n.s.)	-.014 (n.s.)
Score d'immersion total	.072 (n.s.)	.044 (n.s.)

De même, ni le rythme cardiaque moyen brut, ni le rythme cardiaque corrigé n'a permis de discriminer significativement le niveau immersion des participants entre les types d'interfaces ou entre les présentations vidéo (Table 39).

Tableau 39 – Analyse de la variance des indicateurs PPG en fonction du type d'interface et du style de présentation

Sources et mesures	**dll**	**F**	η^2_p
Rythme cardiaque moyen			
Type d'interface (I)	2	2,761	-
Style de présentation (P)	1	0,091	-
I x P	2	0,737	-
Erreur	63		
Rythme cardiaque moyen relatif			
Type d'interface (I)	2	1,376	-
Style de présentation (P)	1	0,397	-
I x P	2	2,801	-
Erreur	63		

Mouvements oculaires

Les analyses statistiques sur les mouvements oculaires ont été réalisées sur 94 participants. Pour le nombre de fixations, on constate une corrélation faible avec le score d'immersion total (r = .183, p = .009) et avec la sous-échelle de plaisir accru (r = .234, p = .001). Pour la durée moyenne par fixation, on constate seulement une corrélation négative faible avec la sous-échelle de plaisir accru (r = -.142, p = .044) ; (Tableau 40).

Tableau 40 – Corrélations entre les mesures oculaires et les échelles du questionnaire d'immersion

	Nombre de fixations	Durée moyenne par fixation
Plaisir accru	**.234****	**-.142***
Immersion concentrée	.128 (n.s.)	-.041 (n.s.)
Dissociation temporelle	.076 (n.s.)	.053 (n.s.)
Score d'immersion total	**.183****	-.057 (n.s.)

*p<.05 **p < .01

Une analyse de la variance (Tableau 41), nous montre que le nombre de fixations et la durée moyenne de fixation ont permis de discriminer significativement entre les types d'interfaces et les deux styles de présentation.

Tableau 41 – Analyse de la variance des mesures oculaires en fonction du type d'interface et du style de présentation

Sources et mesures	dll	F	η^2_p
Nombre de fixations			
Type d'interface (I)	2	**10,223****	.183
Style de présentation (P)	1	**37,755****	.293
I x P	2	0,666	-
Erreur	91		
Durée moyenne par fixation			
Type d'interface (I)	2	**6,491***	.125
Style de présentation (P)	1	**17,942****	.165
I x P	2	0,601	-
Erreur	91		

*p > .01, **p > .001

La figure 124 illustre la différence de fixation oculaire en fonction de la présentation et du type interface. On constate que le nombre total de fixation est plus important pour les conditions immersives, c'est à dire pour la bonne présentation et la bonne interface.

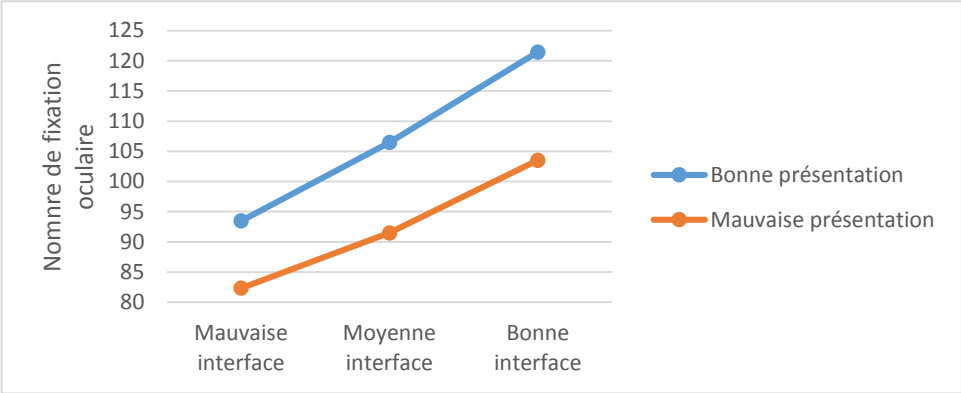


Figure 124 – Nombre de fixations en fonction du type d'interface et du style de présentation

Pour l'évolution du nombre de fixations, on constate une corrélation faible avec le score d'immersion total (r = .183, p = .009) et avec toutes ses sous-échelles. Pour l'évolution de la durée moyenne par fixation, aucune corrélation significative n'a été constatée (Tableau 42).

Tableau 42 – Corrélations entre les mesures d'évolution du comportement oculaire et les échelles du questionnaire d'immersion

	Evolution du nombre de fixations	Evolution de la durée moyenne par fixation
Plaisir accru	**.170****	-.136 (n.s.)
Immersion concentrée	**.151***	-.070 (n.s.)
Dissociation temporelle	**.200***	-.044 (n.s.)
Score d'immersion total	**.183****	-.104 (n.s.)

*p<.05 **p < .01

Une analyse de la variance (Tableau 43), nous montre que l'évolution du nombre de fixations et de la durée moyenne de fixation ont permis de discriminer significativement entre les deux styles de présentation.

Tableau 43 – Analyse de la variance des mesures d'évolution du comportement oculaire en fonction du type d'interface et du style de présentation

Sources et mesures	dll	F	η^2_p
Evolution du nombre de fixations			
Type d'interface (I)	2	1,463	-
Style de présentation (P)	1	**7,929****	.080
I x P	2	0,063	-
Erreur	91		
Evolution de la durée moyenne par fixation			
Type d'interface (I)	2	0,624	-
Style de présentation (P)	1	**10,556****	.104
I x P	2	1,048	-
Erreur	91		

**p > .01

La figure 125 et 126 illustre l'évolution des fixations oculaires au cours du temps en fonction de la présentation et du type interface. On constate que le nombre de fixations diminue moins au cours du temps pour la bonne présentation et la bonne interface que pour les conditions plus mauvaises.

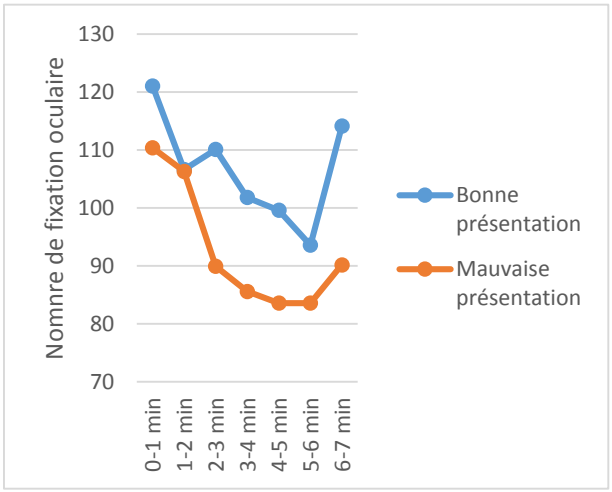


Figure 125 – Evolution des fixations en fonction du style de présentation

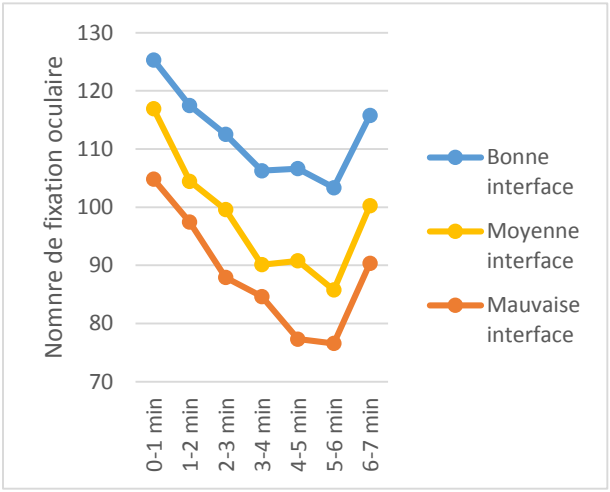


Figure 126 – Evolution des fixations en fonction du type d'interface

Analyse multifacette et multi-mesure (Expérience 1)

Mesures utilisées

L'objectif de cette étude a été de tester, à travers un protocole psychométrique multi-facettes, la fiabilité de certains indicateurs de l'expérience 1 jugés pertinents pour mesurer l'immersion, ainsi que certaines propositions de mesures composites, issues de la combinaison de ces indicateurs. Ces indicateurs ont été renommés pour marquer leurs modalités d'acquisition.

Parmi les indicateurs simples, la mesure auto-rapportée d'immersion, le jugement vidéo de l'immersion et le jugement des traces écrites du plaisir ont été sélectionnés pour l'étude. La mesure auto-rapportée d'immersion est issue du score total obtenu par la moyenne des 9 items du questionnaire d'immersion. Cet indicateur a été jugé pertinent car, en plus d'une bonne consistance interne (α = .852), il a été capable de discriminer significativement entre la « bonne » et la « mauvaise » présentation (F(1,57) = 37,419, p > .000, η^2_p = .396) et a permis également de détecter un effet d'interaction entre le style de présentation et le type d'interface (F(2,57) = 6,038, p = .004, η^2_p = .175). Dans cette étude, cet indicateur est appelé **indicateur subjectif**. Le jugement vidéo de l'immersion est issu de l'évaluation experte du niveau d'immersion des utilisateurs à partir de leurs expressions faciales et corporelles. Cet indicateur a été jugé pertinent car il possède une consistance inter-juge correcte (CCI = .518) et possède la meilleure corrélation avec le score d'immersion total issu du questionnaire d'immersion (r = .540, p > .001). De plus, le jugement vidéo de l'immersion (sans pondération) a été capable de discriminer significativement entre les types d'interfaces (F(1,42) = 3,553, p = .047, η^2_p = .254). Dans cette étude, cet indicateur est appelé **indicateur vidéo.** Le jugement des traces écrites du plaisir est issu de l'évaluation experte du niveau de plaisir des utilisateurs à partir de leurs verbalisations écrites. Cet indicateur a été jugé pertinent car il possède une consistance inter-juge correcte (CCI = .640) et possède la corrélation la plus importante avec le score d'immersion total issu du questionnaire d'immersion (r = .382, p > .01). De plus, le jugement des traces écrites du plaisir a été capable de discriminer significativement entre les styles de présentation (F(1,28) = 14,647, p = .001, η^2_p = .343). Dans cette étude, cet indicateur est appelé **indicateur textuel**. L'indicateur de conductance de la peau n'a pas été retenu. En effet, même si l'on constate une corrélation significative entre la conductance de la peau et le score d'immersion total (r = -.290, p < .01), ainsi qu'une discrimination significative entre les types d'interfaces (F(1,42) = 3,553, p = .04, η^2_p = .145), il est difficile d'interpréter ces résultats en terme d'augmentation de l'immersion, car les conditions avec le module chat sont celles où les utilisateurs ont été amenés à utiliser le clavier, ce qui a pu parasiter la capture de la conductance de la peau. Tous ces indicateurs ont été standardisés pour obtenir un score compris entre 0 et 1.

En complément, deux indicateurs composites ont été construits : un indicateur « *objectif* » et un indicateur composite général. L'**indicateur objectif** est composé de la moyenne de l'indicateur textuel et vidéo. Il est « *objectif* » dans le sens où il s'appuie sur des données non subjectives,

telles que celles recueillies par le questionnaire[92]. Enfin, l'**indicateur composite général** est composé de la moyenne de l'indicateur subjectif et objectif.

Plans d'étude pour l'analyse de la généralisabilité

Un plan d'étude a été mis en place afin d'estimer la fiabilité du protocole de test selon la théorie de la généralisabilité. Il a consisté à identifier les facettes d'intérêts, la nature de leurs relations et leurs types d'échantillonnage. Puis, un objet d'étude a été déterminé, ainsi que le type de mesure visé par le protocole de mesure développé. Enfin, une étude G et plusieurs études D ont été menées pour déterminer le coefficient de généralisabilité absolue du protocole de test en faisant varier le type de mesure sélectionné et le nombre d'utilisateurs.

Dans notre étude, les facettes « *Présentation vidéo* », « *Utilisateur* » et « Indicateur » ont été sélectionnées. La facette « *Présentation vidéo* » représente l'ensemble des vidéos de conférence à distance dont nous cherchons à discriminer les éléments au travers de la notion d'immersion. La facette « Utilisateur » contient tous les participants au test et dont on dispose des trois indicateurs : subjectif, vidéo et textuel (15 utilisateurs). Enfin, la facette « Indicateur » contient les différents indicateurs sélectionnés pour cette étude. Toutes les facettes ont été croisées ensemble, selon un plan d'observation : V x U x I (Figure 127). En effet, tous les utilisateurs ont regardé l'ensemble des vidéos et tous les indicateurs ont été recueillis pour chacun.

Tableau 44 – Univers d'échantillonnage et niveau des facettes

Abrév.	Nom de la facette	Niveau	Univers
V	Présentation Vidéo	5	Infini
U	Utilisateur	30	Infini
I	Indicateur	3	3

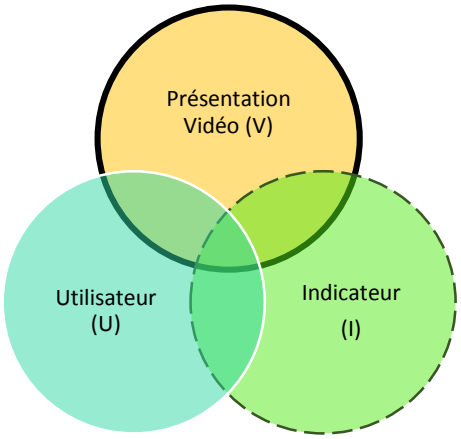


Figure 127 – Plan d'observation VUI

Le protocole prend en compte une facette de différenciation (« *Présentation Vidéo* » à deux modalités) et deux facettes d'erreurs (« *Utilisateur* » à 15 modalités, et « *Indicateur* » à 3 modalités). Les univers d'échantillonnage pour les facettes « *Présentation Vidéo* », et « *Utilisateur* » ont été spécifiés comme infinis (Tableau 44). Il s'agira donc de tester la fiabilité d'un protocole d'évaluation dont les éléments qui le composent, c'est-à-dire les présentations et les participants sollicités, varient d'un test à l'autre. Par contre, l'univers d'échantillonnage de la facette « *Indicateur* » a été fixé à trois. Il s'agira donc de restreindre l'étude des indicateurs de mesure de l'immersion à ceux utilisés dans le test, c'est-à-dire aux indicateurs subjectif, vidéo, textuel et leurs combinaisons.

[92] Néanmoins, il est vrai que le terme objectif peut induire en erreur, dans le sens que cet indicateur s'appuie sur des jugements d'experts, qui sont empreints d'une certaine subjectivité, même s'il est possible dans le futur que l'analyse de ces données vidéo et textuelles soit réalisé automatiquement par un algorithme spécialisé.

Résultats

Le tableau 45 expose l'analyse de la variance et sa décomposition en divers éléments. Il montre que la variation engendrée par la facette de différenciation « *Présentation vidéo* » est particulièrement importante (33%), tout comme la variabilité résiduelle, non expliquée (30,2%). L'interaction entre la facette « *Utilisateur* » et la facette « *Présentation vidéo* » est de 21,7%, ce qui en fait le levier le plus intéressant pour améliorer la qualité de la mesure. La variation pure engendrée par les facettes « *Indicateur* » (0%) et « *Utilisateur* » (0,7%) est quasi-nulle, notamment car l'agrégation des avis des experts, ainsi que les items a permis de « lisser » cette variabilité, qui n'est pas apparente dans ce modèle simplifié.

Tableau 45 – ANOVA et calcul des composants de la variance

				Composants				
Source de variations	**Somme des carrés**	**ddl**	**Carré moyen**	**Aléatoire**	**Mixte**	**Corrigé**	**%**	**Erreur standard**
V	21.74642	1	21.74642	0.40280	0.46293	0.46293	33.0	0.39770
I	4.89326	2	2.44663	-0.02420	-0.02420	-0.01613	0.0	0.09379
U	13.58513	14	0.97037	0.00213	0.00928	0.00928	0.7	0.08313
VI	6.25935	2	3.12967	0.18038	0.18038	0.18038	12.9	0.14771
VU	12.80555	14	0.91468	0.16357	0.30489	0.30489	21.7	0.11380
IU	13.07197	28	0.46686	0.02144	0.02144	0.02144	1.5	0.08141
VIU	11.87108	28	0.42397	0.42397	0.42397	0.42397	30.2	0.10947
Total	84.23276	89					100%	

En se basant sur les données du tableau précédent, le tableau 46 présente les résultats de l'étude G pour le plan V/UI. Le coefficient de généralisabilité absolu obtenu est très élevé ($\Phi = .95$), ce qui traduit une bonne fiabilité du dispositif à déterminer la position exacte de chaque présentation vidéo sur une échelle de mesure de l'immersion (de 0 à 1). Ce bon score s'explique en partie par le nombre satisfaisant d'utilisateurs ($n = 15$) en rapport à sa contribution en termes d'erreur de mesure. Le deuxième facteur qui explique ce score est d'avoir fait le choix de fixer la facette indicateur, ce qui permet de retirer sa variance d'erreur au protocole de mesure. Néanmoins cela implique que les praticiens voulant utiliser ce protocole de mesure, doivent impérativement utiliser dans chacun de leur test les mêmes éléments inclus dans la facette indicateur, c'est-à-dire ici les indicateurs subjectif, vidéo et textuel. Enfin, le facteur le plus important pour expliquer ce score est la grande part de variance portée par la facette de différenciation. En effet, la variance d'erreur étant beaucoup plus faible que la variance de différenciation, le protocole de mesure est arrivé sans mal à discriminer les présentations vidéo à partir des indicateurs et des utilisateurs disponibles.

Tableau 46 – Répartition de la variance et calcul des coefficients de généralisabilité

Variance de différenciation		Variance d'erreur				
Source	**Variance**	**Source**	**Variance d'erreur relative**	**%**	**Variance d'erreur Absolue**	**%**
V	0.43992		.....		.....	
	.....	I	.....		(0.00000)	0.0
	.....	U	.....		0.00093	4.1
	.....	VI	(0.00000)	0.0	(0.00000)	0.0
	.....	VU	0.02201	100.0	0.02201	95.9
	.....	IU	.....		(0.00000)	0.0
	.....	VIU	(0.00000)	0.0	(0.00000)	0.0
Somme de la variance	0.43992		0.02201	100%	0.02294	100%
Écart-type	0.66326		**Écart-type relatif :** 0.14834		**Écart-type absolu :** 0.15146	
Ep^2	0.95					
Φ	0.95					

En manipulant la facette « Indicateur » et « utilisateur », les études D menées nous permettent de comparer le coefficient de généralisabilité absolu en fonction des indicateurs de mesure de l'immersion retenus et combinés pour ce protocole de test (Figure 128).

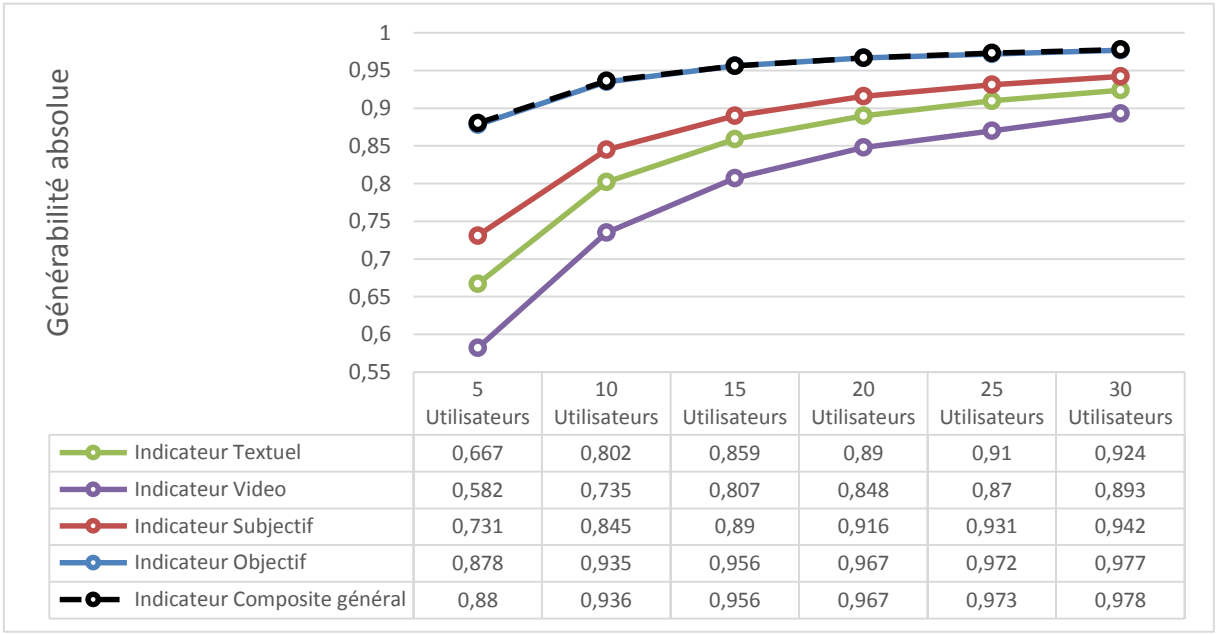


Figure 128 - Etude D sur l'évolution du coefficient de généralisabilité absolue en fonction du nombre d'utilisateur et du type d'indicateur de mesure de l'immersion

On constate ainsi que l'indicateur subjectif présente un coefficient de généralisabilité absolu supérieur à celui de l'indicateur vidéo et textuel, quel que soit le nombre d'utilisateur participant au test. Ainsi, pour discriminer de manière fiable les présentations vidéo en fonction de l'immersion de l'utilisateur, il faut au minimum 10 utilisateurs pour l'indicateur subjectif (Φ = 0,845), 15 utilisateurs pour l'indicateur textuel (Φ = 0,859) et 20 utilisateurs pour l'indicateur vidéo (Φ = 0,848). L'indicateur composite « Objectif », qui combine les données des indicateurs

vidéo et textuel, présente des performances très élevés. En effet, il obtient un coefficient de généralisabilité absolu de 0,956 pour 15 utilisateurs, ce qui montre une fois encore l'avantage d'agréger des indicateurs de fiabilité proches mais de modalités très différentes. La qualité de cet indicateur rivalise avec l'indicateur composite général, qui agrège les trois modalités, et dont les performances se confondent presque totalement avec l'indicateur composite objectif.

Néanmoins, il convient de rappeler que cette étude G ne se base que sur la facette de différentiation « *présentation vidéo* », contenant deux conditions croisées bien tranchées De ce fait, si les comparaisons relatives entre mesures restent intéressantes, les données chiffrées absolues seront plus représentatives dans l'expérience 2 où la facette de différenciation contient un nombre d'éléments plus important (six éléments contre deux) et issus de conditions croisées et emboitées.

Analyse multifacette et multi-mesure (Expérience 2)

Mesures utilisées

Cette étude a pour objectif de tester, à travers un protocole psychométrique multi-facettes, la fiabilité de certains indicateurs de l'expérience 2 jugés pertinents. Néanmoins, il est apparu lors de l'analyse mono-mesure des indicateurs de l'expérience 2 que seuls les indicateurs subjectifs ont présenté une fiabilité et une validité suffisante. De ce fait, seul l'indicateur subjectif du questionnaire d'immersion a été incorporé à l'étude G, principalement à des fins de comparaison avec l'étude G de l'expérience 1. D'autres indicateurs ont été incorporés dans les études D, pour comparer la fiabilité entre les indicateurs de l'expérience 2. Tous ces indicateurs ont également été renommés pour marquer leurs modalités d'acquisition. Dans cette étude, la mesure auto-rapportée d'immersion, la mesure mono-item de l'immersion, la durée moyenne de fixation et la conductance de la peau relative ont été sélectionnée.

La mesure auto-rapportée d'immersion est issue du score total obtenu par la moyenne des 9 items du questionnaire d'immersion. Cet indicateur a été jugé pertinent car, en plus d'une bonne consistance interne (α = .868), il a été capable de discriminer significativement entre les présentations vidéo ($F(1,100) = 105{,}203$, $p > .000$, $\eta^2_p = .513$) et entre les types d'interfaces ($F(1,100) = 6{,}798$, $p = .002$, $\eta^2_p = .120$). Dans cette étude, cet indicateur est appelé **indicateur 9-items**. La mesure mono-item de l'immersion est issue de l'échelle ordinale en 10 points demandant à l'utilisateur de situer son niveau d'immersion lors de l'expérience. Cet indicateur a été jugé pertinent car il présente une corrélation très forte avec le score total du questionnaire d'immersion (r = .786) et a permis de discriminer significativement entre la « bonne » et la « mauvaise » présentation ($F(1,100) = 153{,}942$, $p > .000$, $\eta^2_p = .606$) et entre les types d'interfaces ($F(1,100) = 13{,}048$, $p > .000$, $\eta^2_p = .207$), soit avec une sensibilité de discrimination supérieure à celle observée pour l'indicateur 9 item. Dans cette étude, cet indicateur est appelé **indicateur 1-item.**

Deux autres indicateurs, qui se sont montrés peu pertinents lors de l'analyse psychométrique mono-mesure, ont été ajoutés à des fins de comparaison pour les études D. Il s'agit de la durée moyenne de fixation et de la conductance de la peau relative. La durée moyenne de fixation a été choisie car elle est plus générique que le nombre de fixations, pouvant fortement varier

d'une tâche à l'autre. Cet indicateur a présenté une corrélation négative faible avec la sous-échelle de plaisir accru (r = -.142, p = .044) et a permis de discriminer significativement entre les types d'interface (F(2,91) = 6,491, p > .01, η^2_p = .125) et les deux styles de présentation (F(1,91) = 17,942 > .001, η^2_p = .165). Néanmoins, en termes de validité, il n'est pas encore possible de tirer des conclusions solides car les liens observés entre cette indicateur et l'immersion vont dans le sens contraire de la littérature (Jennett et al., 2008). Dans cette étude, cet indicateur est appelé **indicateur oculaire.** La conductance de la peau relative a été également choisie à des fins de comparaison. Cet indicateur a présenté une corrélation négative faible avec le score d'immersion total (r = -.145, p = .046), a permis de discriminer significativement entre les types d'interface (F(2,93) = 6,325, p = .003, η^2_p = .145) et les présentations (F(1,93) = 6,102, p = .015, η^2_p = .062). Néanmoins, cet indicateur a montré une indexation non linéaire avec l'immersion, car cette dernière varie dans un sens pour les types d'interfaces et dans l'autre pour les présentations, ce qui pose un sérieux problème si l'on souhaite l'utiliser en tant que mesure de l'immersion. Dans cette étude, cet indicateur est appelé **indicateur cutané.**

Aucun indicateur composite n'a été construit dans cette étude car il n'y avait pas d'indicateurs de modalités différentes possédant des niveaux de fiabilité proche.

Plans d'étude pour l'analyse de la généralisabilité

Un plan d'étude a été mis en place afin d'estimer la fiabilité du protocole du test selon la théorie de la généralisabilité. Une étude G et plusieurs études D ont été menées pour déterminer le coefficient de généralisabilité absolu du protocole de test en faisant varier le type de mesures sélectionnées et le nombre d'utilisateurs.

Dans notre étude, les facettes « *Présentation vidéo* », « Type d'interface », « *Utilisateur* » et « Indicateur » ont été sélectionnées. Les facettes « *Présentation vidéo* » et «*Type d'interface »* représentent l'ensemble des vidéos et interfaces de conférence à distance dont nous cherchons à discriminer les éléments au travers de la notion d'immersion. La facette « *Utilisateur* » contient tous les participants au test et dont on dispose des trois indicateurs : subjectif, oculaire et cutanée (81 utilisateurs ou 3x27, si on les répartit dans les trois conditions d'interface). Enfin, la facette « *Indicateur* » contient les différents indicateurs sélectionnés pour cette étude. La facette « *Utilisateur* » est nichée dans la facette « Type d'interface », car les utilisateurs ont interagit avec seulement une des trois interfaces lors de l'expérience. Les autres facettes ont été croisées car tous les utilisateurs ont regardé l'ensemble des vidéos et tous les indicateurs ont été recueillis pour chacun. Le plan d'observation pour cette étude est donc : (U : T) x V x I (Figure 129).

Le protocole prend en compte deux facettes de différenciation (« Présentation Vidéo » à deux modalités, et « *Type d'interface* » à trois modalités) et deux facettes d'erreurs (« *Utilisateur* » à 27 modalités, et « *Indicateur* » à 4 modalités). Les univers d'échantillonnage pour les facettes « *Présentation Vidéo* », et « *Utilisateur* » ont été spécifiés comme infini (Tableau 47). Il s'agira donc de tester la fiabilité d'un protocole d'évaluation dont les éléments qui le composent, c'est-à-dire les présentations, les interfaces et les participants sollicités, varient d'un test à l'autre. Par contre, l'univers d'échantillonnage de la facette « *Indicateur* » a été fixé à quatre. L'étude sera donc restreinte aux indicateurs de mesure de l'immersion utilisés dans le test.

Tableau 47 – Univers d'échantillonnage et niveau des facettes

Abrév.	Nom de la facette	Niveau	Univers
V	Présentation Vidéo	5	Infini
T	Type d'interface	3	Infini
U:T	Utilisateur	27	Infini
I	Indicateur	4	4

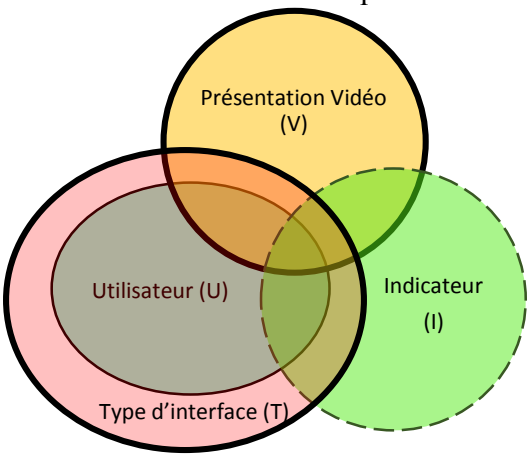


Figure 129 – Plan d'observation (U:T)VI

Résultats

Le tableau 48 expose l'analyse de la variance et sa décomposition en divers éléments en ne gardant que l'indicateur d'immersion du questionnaire en 9 items (Cardinet, Johnson, & Pini, 2011, p. 48). Il montre que la variation engendrée par la facette de différentiation « *Présentation vidéo* » est particulièrement importante (35,6%), tout comme la variabilité résiduelle, non expliquée (35,5%). Ce résultats est très comparable à celui de l'étude G de l'expérience 1 où la variation engendrée par la facette de différenciation « *Présentation vidéo* » était de 33% et la variabilité résiduelle de 30,2%.

Tableau 48 – ANOVA et calcul des composants de la variance avec restriction de la facette « indicateur » au seul niveau « indicateur 9-items »

Source de variations	Somme des carrés	ddl	Carré moyen	Composants				
				Aléatoire	Mixte	Corrigé	%	Erreur standard
V	1.26405	1	1.26405	0.01560	0.01560	0.01560	35.6	0.01274
T	0.25879	2	0.12940	0.00199	0.00199	0.00199	4.5	0.00170
I	.....	.....	.....	.....	.....	.....	.....	.....
U:T	2.88094	78	0.03694	0.01068	0.01068	0.01068	24.4	0.00317
VT	0.00127	2	0.00064	-0.00055	-0.00055	-0.00055	0.0	0.00009
VI	.....	.....	.....	.....	.....	.....	.....	.....
VU:T	1.21433	78	0.01557	0.01557	0.01557	0.01557	35.5	0.00246
UI	.....	.....	.....	.....	.....	.....	.....	.....
IU:T	.....	.....	.....	.....	.....	.....	.....	.....
VTI	.....	.....	.....	.....	.....	.....	.....	.....
VIU:T	.....	.....	.....	.....	.....	.....	.....	.....
Total	5.61938	161					100%	

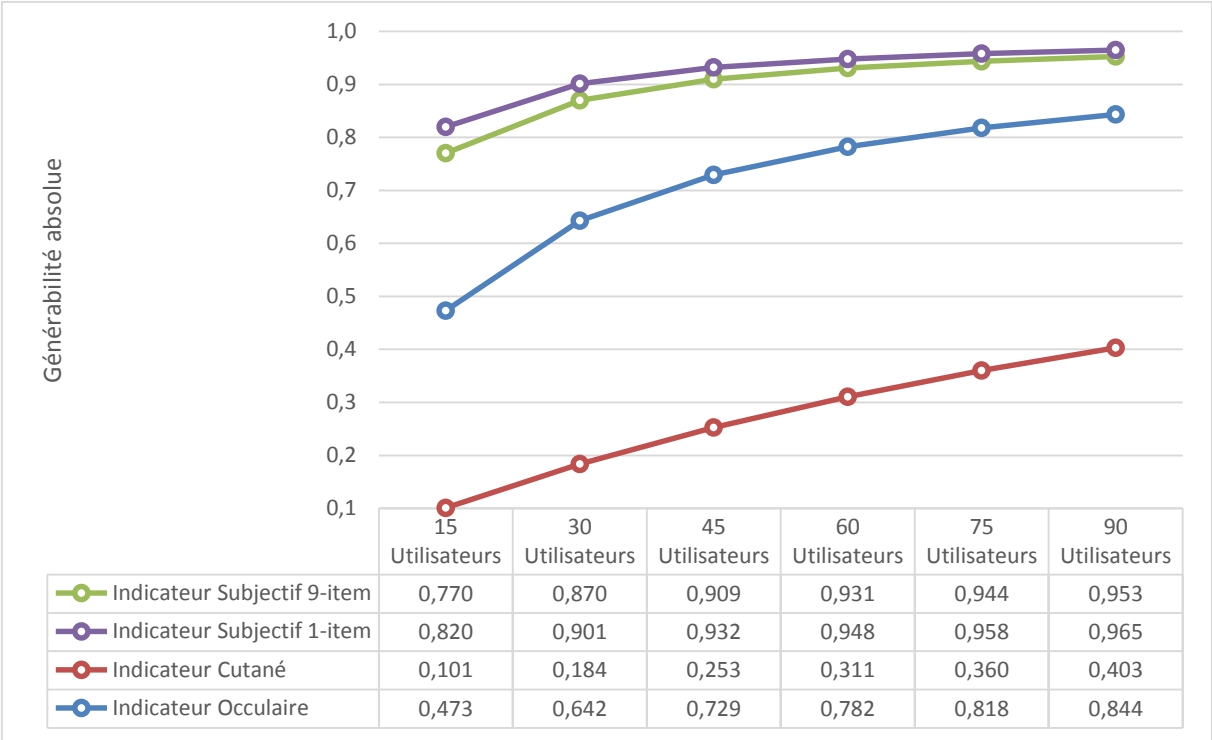
De même, la variation engendrée par la facette « Utilisateur », nichée dans la facette « interface », est également importante (24,4%), ce qui en fait un levier également intéressant pour améliorer la qualité de la mesure. Néanmoins, la différence notable avec l'expérience 1 est la faible variation engendrée par la facette de différenciation « *type d'interface* » (4,5%). Cela induit une plus grande difficulté pour le protocole de mesure à être fiable car cela signifie que les écarts entre les éléments à discriminer sont petits par rapport à la fluctuation aléatoire générée par les facettes de mesure. Cela tranche avec l'expérience 1 où la seule facette de différenciation était portée avec la condition croisée « *présentation vidéo* », qui possédait une grande variance de différenciation.

Tableau 49 – Répartition de la variance et calcul des coefficients de généralisabilité avec restriction de la facette « indicateur » au seul niveau « indicateur 9-items »

Variance de différenciation		**Variance d'erreur**				
Source	**Variance**	**Source**	**Variance d'erreur relative**	**%**	**Variance d'erreur Absolue**	**%**
V	0.01560		.....		.....	
T	0.00199		.....		.....	
	.....	I	.....		.....	
	.....	U:T	0.00040	40.7	0.00040	40.7
VI	(0.00000)		.....		.....	
	.....	VI	.....		.....	
	.....	VU:T	0.00058	59.3	0.00058	59.3
	.....	TI	.....		.....	
	.....	IU:T	.....		.....	
	.....	VII	.....		.....	
	.....	VIU:T	.....		.....	
Somme de la variance	0.01759		0.00097	100%	0.00097	100%
Écart-type	0.13261		**Écart-type relatif :** 0.03118		**Écart-type absolu :** 0.03118	
Ep^2	0.95					
Φ	0.95					

En se basant sur les données du tableau précédent, le tableau 49 présente les résultats de l'étude G pour le plan V/UI. Le coefficient de généralisabilité absolu obtenu est très élevé (Φ = .95), ce qui traduit une bonne fiabilité du dispositif à déterminer la position exacte de chaque présentation vidéo sur une échelle de mesure de l'immersion (de 0 à 1). Ce bon score s'explique grandement par le nombre important d'utilisateurs (81) en rapport à sa contribution en termes d'erreur de mesure. A titre de comparaison, dans l'expérience 1, le questionnaire d'immersion avait obtenu un coefficient de généralisabilité absolu similaire pour un échantillon estimé de 30 utilisateurs. En effet, la variance de différenciation étant plus faible, le protocole de mesure a besoin de plus d'utilisateurs pour arriver au même niveau de précision. En effet, il est plus facile de mettre au point un protocole de test fiable dont le but est de tester l'immersion de plusieurs vidéoconférences visionnées par tous (condition croisée), qu'un protocole de mesure où cet exercice de discrimination prend en compte l'interface de visionnage, testé seulement par une partie de l'échantillon utilisateur (condition emboitée).

En manipulant la facette « *Indicateur* » et « *Utilisateur* », les études D menées nous permettent de comparer le coefficient de généralisabilité absolu en fonction des indicateurs sélectionnés pour ce protocole de test (Figure 130). On constate ainsi que l'indicateur subjectif 1-item présente un coefficient de généralisabilité absolu supérieur à tous les autres indicateurs, quel que soit le nombre d'utilisateurs. En effet, ce dernier est acceptable à partir de 15 utilisateurs (Φ = 0,820) et excellent à partir de 30 (Φ = 0,901). L'indicateur subjectif 9-items présente également de très bon résultats (Φ = 0,870 pour 30 utilisateurs), bien que moindre par rapport à ceux de l'expérience 1, à cause de facettes de discrimination plus exigeantes en terme de mesure. L'indicateur oculaire présente un coefficient de généralisabilité absolu acceptable qu'à partir de 75 utilisateurs (Φ = 0,818). Enfin, l'indicateur cutané est capable de discriminer de manière satisfaisante l'immersion entre les éléments des facettes (interfaces et vidéos) qu'à partir de 600 utilisateurs (Φ = 0,818).



	15 Utilisateurs	30 Utilisateurs	45 Utilisateurs	60 Utilisateurs	75 Utilisateurs	90 Utilisateurs
Indicateur Subjectif 9-item	0,770	0,870	0,909	0,931	0,944	0,953
Indicateur Subjectif 1-item	0,820	0,901	0,932	0,948	0,958	0,965
Indicateur Cutané	0,101	0,184	0,253	0,311	0,360	0,403
Indicateur Occulaire	0,473	0,642	0,729	0,782	0,818	0,844

Figure 130 – Etude D sur l'évolution du coefficient de généralisabilité absolue en fonction du nombre d'utilisateur et du type d'indicateur de mesure de l'immersion

Discussions et Conclusion

En incorporant un grand nombre d'indicateurs de l'immersion, cette étude nous offre un grand nombre d'informations sur leur pertinence et leur fiabilité relatives. Cela concerne autant les indicateurs simples que les indicateurs composites. Cette étude nous permet également d'avoir un retour critique sur la pertinence de la procédure de validation multi-facettes dans le domaine des IHMs.

La mesure auto-rapportée d'immersion, au travers du questionnaire en 9 items, a montré de bonnes performances psychométriques, que ce soit dans la première ou la deuxième expérience. Dans la première expérience, elle a montré un bon niveau de consistance interne (α = .852), une bonne sensibilité pour trancher entre la « bonne » et la « mauvaise » présentation (η^2_p =

.396), et a permis de détecter un effet d'interaction entre le style de présentation et le type d'interface (η^2_p = .175). Une étude de généralisabilité complémentaire nous montre que seulement 10 utilisateurs sont nécessaires sous ces conditions pour disposer d'un score d'immersion fiable permettant de distinguer entre les présentations vidéos (Φ = 0,845). Ces résultats ont été confirmés dans l'expérience 2, grâce à une consistance interne d'un niveau similaire (α = .868) et une capacité à discriminer finement entre les présentations (η^2_p = .513) et les types d'interfaces (η^2_p = .120). Dans ce contexte, l'étude de généralisabilité nous montre que seulement 30 utilisateurs sont nécessaires pour disposer d'un score d'immersion fiable permettant de distinguer des situations variables en terme d'interface et de présentations vidéos (Φ = 0,820). Tous ces résultats nous confirment que le questionnaire d'immersion, même réduit à 9 items, est un outil fiable pour mesurer et étudier l'immersion.

L'indicateur subjectif à 1 item a montré dans cette étude des résultats prometteurs. En effet, il a, d'une part, présenté une corrélation très forte avec le score total du questionnaire d'immersion (*r* = .786), et, d'autre part, fait preuve d'une sensibilité de mesure remarquable, en discriminant très finement les niveaux d'immersion entre les présentations (η^2_p = .606) et types d'interfaces (η^2_p = .207). Cela va dans le sens de la tendance de ces dernières années (Schweizer, 2011) qui met en avant la pertinence d'utilisation de questionnaires très courts (ici, un seul item). Plusieurs hypothèses pourraient expliquer ces résultats. Une première explication pourrait venir du contenu du questionnaire long dont certains items auraient nui à sa fiabilité. Or, l'évaluation de l'impact du retrait individuel de chacun des 9 items, montre que, dans aucun cas, la consistance interne du questionnaire (qui est déjà très bonne) n'augmente significativement (+0,022 au maximum). La deuxième hypothèse pour expliquer la sensibilité plus importante de l'indicateur subjectif en 1 item serait l'utilisation d'une échelle en 10 points à la place de 5. En effet, il est possible que les utilisateurs aient ainsi pu exprimer un jugement de leurs immersions de manière plus fine à partir d'une échelle en 10 points, ce qui aurait pu influer ainsi sur la sensibilité de la mesure. Enfin, la dernière hypothèse est que cette échelle simple, demandant simplement à l'utilisateur de situer son niveau d'immersion sur une échelle ordinale en 10 points, ai bénéficié indirectement du questionnaire en 9 items, par effet de halo. En effet, le remplissage préalable du questionnaire long aurait pu permettre de cerner précisément les différentes dimensions de l'immersion et de les restituer sous la forme d'une note de synthèse dans l'échelle en 10 points.

Le jugement vidéo de l'immersion a montré un bon potentiel pour évaluer l'immersion des participants à partir des enregistrements vidéo de webcam. En effet, il a montré une forte corrélation avec le score d'immersion total (*r* = .540) et a été est capable de discriminer significativement entre les types d'interfaces (η^2_p = .254). Dans l'analyse multifacettes des données issues de l'expérience 1, il a montré également qu'une vingtaine d'utilisateurs suffit pour discriminer l'immersion de manière fiable entre les présentations vidéo (Φ = 0,848, n = 20). Ces résultats montrent la valeur des signaux non-verbaux pour nous renseigner de l'état d'immersion d'un utilisateur. Néanmoins, si pour le moment, de tels signaux complexes peuvent être sans trop de difficultés captés par un petit groupe d'humains (CCI = .518, n = 5), les algorithmes actuels ont encore du chemin à faire pour arriver à de tels résultats. Des études complémentaires sur les signaux pertinents, leurs intégrations et leurs pondérations seront à mener, afin de nourrir ces algorithmes de reconnaissance. Les quelques données qualitatives récoltées auprès des experts ayant participé à l'étude (Annexe 3D), nous montrent à quel point

ces recherches peuvent enrichir nos connaissances sur le domaine. De l'autre côté, le jugement vidéo du plaisir n'a pas donné de résultats significatifs. En effet, on ne constate pas de corrélations significatives avec le score d'immersion total ou ses sous-échelles, et il ne permet pas non plus de discerner significativement entre les présentations vidéo. Il est intéressant de noter que les premières tentatives d'exploitation des expressions faciales dans les milieux des IHM ont été réalisées sous l'angle des émotions, et plus précisément sous l'angle des émotions primaires. *Facereader*, par exemple, analyse les expressions faciales à partir de 6 émotions de base : la joie, la tristesse, la peur, le dégout, la surprise et la colère (den Uyl et al., 2005; Bieke Zaman & Shrimpton-smith, 2006). Or, rien n'indique que ces états subjectifs ne soient, ni les plus fréquents, ni les plus pertinents, à analyser dans le cadre d'interactions avec un système numérique. En effet, d'autres états subjectifs, qui commencent à se faire une place dans la littérature de la mesure d'expérience utilisateur, mériteraient d'être analysés aussi sous l'angle des manifestations non verbales, telles de l'attention soutenue, le flow, l'immersion ou, au contraire, la frustration ou l'ennui. Enfin, la pondération par l'indice de confiance s'est montrée inutile, voire contreproductif, car ces indicateurs ont montré des performances similaires, voire moindres, que leurs équivalents sans pondération. Cela peut s'expliquer par le fait que les juges ont eu les mêmes difficultés sur les mêmes vidéos, et que donc, il n'y a pas eu d'impact positif en moyenne pour une pondération par indice de confiance. En effet, on constate que les indices de confiance donnés pour chacun des jugements sont assez similaires d'un juge à l'autre : CCI = .644 pour le plaisir et CCI = 601 pour l'immersion.

Le jugement des traces écrites du plaisir a également montré un bon potentiel pour évaluer l'immersion des participants à partir des échanges recueillis sur le module de chat. En effet, il a montré une forte corrélation avec le score d'immersion total (r = .382) et a été capable de discriminer significativement entre les types d'interface (η^2_p = .343). Dans l'analyse multifacettes des données issues de l'expérience 1, il montre également qu'une quinzaine d'utilisateurs suffit (Φ = 0,848, n = 15) pour discriminer l'immersion de manière fiable entre les présentations vidéo. Ces résultats nous montrent la valeur des analyses textuelles, qui, par l'intermédiaire de l'évaluation du plaisir/déplaisir, peut nous renseigner sur l'état d'immersion d'un utilisateur. Néanmoins, le passage d'une évaluation experte à une évaluation automatique risque d'être encore plus périlleuse qui pour les expressions faciales et corporelles, dans le sens où la compréhension du niveau émotionnel dans le langage tient compte d'un grand nombre de facteurs, qui ne se limitent pas à la valence des mots, mais également à la formulation de désaccords, de la compréhension de l'ironie, etc. A ce sujet, les quelques données qualitatives compilées dans l'annexe 3D montrent que les jugements effectués vont beaucoup plus loin que les algorithmes d'analyse textuelle de sentiments actuels (De Choudhury et al., 2012; Neviarouskaya et al., 2010; Zhang & Yap, 2012). De l'autre côté, le jugement des traces écrites de l'immersion ont montré des résultats plus mitigés. En effet, même si l'on constate une corrélation significative avec le score d'immersion total (r = .381), ce dernier n'a pas permis de discerner significativement entre les présentations vidéo. Ainsi, il a été difficile pour les experts de noter l'immersion à partir des échanges textuels, dans le sens où l'absence d'échanges peut être interprété comme une preuve d'attention soutenue, et, en même temps, des échanges construits peut être vus comme un critère de bonne implication. Ces indices contraires vont tous deux dans le sens de l'engagement, et donc de l'immersion… ce qui pose problème. Enfin, la

pondération par l'indice de confiance s'est montrée également inutile, les juges ayant ressenti de manière consistante les mêmes difficultés et facilités (CCI = .853 pour le plaisir, et CCI = .851 pour l'immersion).

Les données recueillies de conductance de la peau n'ont pas permis de trancher nettement sur son utilité pour mesurer l'immersion. Tout d'abord, pour la conductivité moyenne de la peau et la conductivité de la peau normalisée, aucun lien significatif avec l'échelle auto-rapportée d'immersion et aucune discrimination significative entre les présentations ou les interfaces n'ont été trouvées, que cela soit dans l'expérience 1 ou 2. Pour la conductivité moyenne de la peau, ces résultats ne sont pas surprenants, étant donné la grande différence interindividuelle existant pour le niveau de conductivité cutanée au repos (Dawson et al., 2007). C'est pour cette raison que des indicateurs essayant d'éliminer ces différences interindividuelles ont été mis au point. Cela est plus étonnant pour la conductivité de la peau normalisée, qui devait, à partir d'une mesure supplémentaire du niveau maximum de conductivité cutané, normaliser les écarts en termes de variation maximum entre individus. Une analyse complémentaire de ces données nous montre que le jeu vidéo utilisé en fin de passation a eu du mal à stimuler assez les utilisateurs pour récupérer un niveau de conductivité qui soit significativement supérieur aux niveaux observés durant la passation. Cette étape de *« stimulation maximale* » devra donc être améliorée pour permettre un engagement profond pour tous les utilisateurs. D'un autre côté, l'indicateur de conductivité de la peau relative a donné de meilleurs résultats, bien que difficile à interpréter. En effet, le sens des relations observé a varié d'une expérience à l'autre, voire même dans la même expérience. En premier, on constate une corrélation négative modérée avec le score d'immersion total dans l'expérience 1 (r = -.290) et faible dans l'expérience 2 (r = .145). Cela implique que la conductance de la peau relative a baissé quand le niveau d'immersion a augmenté. D'un autre côté, dans l'expérience 1, cette indicateur a détecté une différence significative entre les types d'interface ($F(1,42) = 3,553$, $p = .04$, $\eta^2_p = .145$), alors qu'aucun autres indicateurs, ayant pourtant démontré de bonnes sensibilités psychométriques, n'en a trouvé. A l'inverse, cet indicateur n'a pas détecté de différence en termes d'immersion entre les présentations vidéo, alors que les autres indicateurs y sont arrivés. Il est donc fort probable que les utilisations du clavier aient pu parasiter la capture de la conductance de la peau dans l'expérience 1, ce qui expliquerait une conductivité de la peur relative supérieur dans les conditions avec le module de chat. Enfin, dans l'expérience 2, cet indicateur a permis de discriminer significativement entre les types d'interface ($\eta^2_p = .145$) et les présentations (η^2_p = .062) mais en variant en sens inverse dans chacun des deux facteurs de différentiation (vidéos et interfaces). En effet, on constate que la conductivité de la peau relative baisse de la condition d'interface la moins immersive à la plus immersive, mais augmente de la condition de présentation la moins immersive à la plus immersive. Pour expliquer ce phénomène, il est possible d'avancer l'hypothèse d'une réponse non linéaire avec l'immersion, qui prendrait la forme d'une courbe en U inversé. Il s'agirait de mettre sur un axe horizontal le niveau de stimulation de la situation et en vertical le niveau d'engagement. A gauche ou à droite le niveau d'engagement n'est pas optimum car la situation est soit trop complexe (trop stimulante), soit

trop simple (pas assez stimulante)[93] ; (Figure 129). Dans cette expérience, l'axe horizontal est représenté par la conductance de la peau, qui est un bon indicateur du niveau d'activation physiologique du participant, et donc son niveau de stimulation cognitive. Ainsi, nous pouvons expliquer l'augmentation de la conductance de la peau de la « *Mauvaise présentation* » à la « B*onne présentation* » en la plaçant du côté gauche de la cloche d'immersion. En effet, si l'on suit cette hypothèse, la mauvaise présentation, à cause de la pauvreté des plans de cadrage et une performance oratoire monotone et ennuyeuse, a pu placer l'utilisateur dans la zone de sous-stimulation. Au contraire, la « Bonne présentation », avec ses plans dynamiques et une performance oratoire plus intéressante, a pu plus approcher l'utilisateur de la zone optimale, rendant possible l'immersion. D'un autre côté, là où la bonne interface n'a pas représenté une entrave à l'immersion, la mauvaise interface a pu placer l'utilisateur dans la zone de sur-stimulation, à droite de la cloche. En effet, à cause d'un cadre vidéo plus petit et de plus mauvaise qualité (et donc plus difficile à suivre), accompagné d'un florilège de distracteurs sonores et visuels, cette condition aurait augmenté la charge cognitive et la frustration des utilisateurs. Si l'on accepte cette hypothèse, cela rendrait difficile l'exploitation de la conductance de la peau en tant que mesure de l'immersion, car cela nécessiterait une connaissance de la position de cette cloche pour chaque utilisateur. Cela nous semble une entreprise délicate car, rien que d'un point de vue physiologique, il nous est apparu difficile d'identifier le niveau maximum et minimum de conductivité de la peau pour chaque individu. Cela nécessite de mettre au point des procédures identiques permettant de relaxer totalement ou d'engager au maximum tous les participants. Il faudrait en plus de cela, détecter le niveau optimum de stimulation pour chaque personne, pour ainsi en déduire une mesure de l'écart à ce maximum, dont l'éloignement en valeur absolu implique une dégradation du niveau d'immersion. Or il a été montré que la localisation de ce pic hédonique peut également varier d'une personne à l'autre (Dorfman & McKenna, 1966), ce qui rend très difficile l'utilisation de cette mesure dans le cadre d'une utilisation sortant d'un environnement contrôlé, mise à part peut-être pour des cas d'usages très spécifiques. Les données recueillies de rythme cardiaque n'ont pas permis de détecter un lien avec l'immersion, que cela soit pour l'indicateur brut ou relatif. En effet, aucun lien significatif avec l'échelle auto-rapportée d'immersion et aucune discrimination significative entre les présentations ou les interfaces n'ont été trouvé. La première hypothèse avancée pour justifier cette absence de résultats est que la moyenne du rythme cardiaque sur une longue session (7 min) n'est pas un indicateur assez sensible pour détecter l'immersion. En effet, les études parlant généralement de baisse ou d'augmentation du rythme cardiaque se réfèrent souvent à une stimulation relativement proche temporellement. Ainsi, rien n'indique que cette réponse ou série de réponses puissent être détectées par une moyenne du rythme cardiaque sur une période aussi longue. La deuxième hypothèse, qui peut se cumuler avec la première, est liée aux causes multiples de variation du rythme cardiaque. En

[93] Ce modèle a été repris au cours du temps, depuis la célèbre courbe de Wundt (Figure 129), en passant par les travaux de Berlyne (1960, 1970) ou le modèle d'adaptabilité de Handcock et Warm (1989). Seuls les libellés des états de l'axe horizontal changent : sous-stimulation et sur-stimulation pour Wundt, familiarité totale et absence totale de familiarité pour Berlyne, et hypostress et hyperstress pour Handcock et Warm. On pourrait également ajouter le modèle du flow, qui reconnait une zone d'apathie/d'ennui d'un côté et d'anxiété/détresse de l'autre côté, dès que l'équilibre entre les compétences de l'utilisateur et le challenge de la tâche ne sont pas atteint. Dans tous les cas, l'état optimal se trouve au sommet de la cloche.

effet, une activation accrue de la branche sympathique du système nerveux autonome, lors de la présentation d'un stimulus sensoriel intense, d'un stress, ou d'une charge émotionnelle, a pour effet d'augmenter le rythme cardiaque. D'un autre coté, l'influence de la branche parasympathique du système nerveux autonome, lors d'un traitement attentif d'un stimulus, fait baisser le rythme cardiaque (Frijda, 1986). Ces deux influences simultanées, font que sur un enregistrement de plusieurs minutes, la mesure de l'immersion, plutôt portée par l'activation branche parasympathique, peut être bruitée par l'influence simultanée de la branche sympathique. Il existe toutefois un indicateur cardiaque qui permet d'isoler uniquement l'influence parasympathique : c'est la composante haute fréquence du spectre de la variabilité du rythme cardiaque (HRV ;Rowe, Sibert, & Irwin, 1998 ; Hansen, Johnsen, & Thayer, 2003). Cet indicateur, s'il est bien adapté à l'analyse d'enregistrement cardiaque long, a pour faiblesse d'être très sensible aux artéfacts de mesure (Kaufmann, Sütterlin, Schulz, & Vögle, 2011). Or, et cela rejoint notre troisième hypothèse sur la cause du manque de résultats, les données cardiaques de cette étude ont enregistré en moyenne plus de 22 artefacts de mesure, sans compter les rejets de 37 utilisateurs, à cause d'un signal trop parasité. De ce point, de vue, même si la capture par la technique de la photopléthysmographie a présenté l'avantage d'être beaucoup moins invasive que l'électrocardiogramme, cela ne nous a pas permis de recueillir des données d'une qualité suffisante pour effectuer ces analyses.

Les données de mouvements oculaires recueillies ne nous ont pas permis de trancher nettement sur leurs pertinences dans le cadre de la mesure de l'immersion. Tout d'abord, on constate une série de liens assez faibles entre ces mesures et les échelles du questionnaire d'immersion. Ainsi, pour le nombre de fixations, on constate une corrélation faible avec le score d'immersion total (r = .183) et avec la sous-échelle de plaisir accru (r = .234). Pour la durée moyenne par fixation, on constate seulement une corrélation négative faible avec la sous-échelle de plaisir accru (r = -.142). Pour l'évolution du nombre de fixations, on constate une corrélation faible avec le score d'immersion total (r = .183) et avec toutes ses sous-échelles (de .151 à .200). Enfin, pour l'évolution de la durée moyenne par fixation, aucune corrélation significative n'a été constatée. D'un autre côté, les données de l'expérience 2 nous montre que ces indicateurs discriminent modérément les différentes conditions de l'expérience. En effet, le nombre de fixations et la durée moyenne de fixation ont permis de discriminer significativement entre les types d'interfaces et les deux styles de présentation. D'un autre côté, et l'évolution du nombre de fixations et de la durée moyenne par fixation ont permis de discriminer significativement entre les styles de présentation seulement. Si ces résultats statistiques sont en soi plutôt encourageant, un problème se pose : la direction de ces tendances. En effet, les liens observés vont dans le sens inverse de ceux obtenus par d'autres auteurs sur le sujet. Par exemple, l'étude menée par Jennett et al. (2008) a montré que le nombre de fixations diminue sur une tâche immersive au cours du temps ; alors que, au contraire, sur une tâche non immersive, le nombre de fixations augmente car l'individu serait plus distrait. Ainsi, ces auteurs en déduisent qu'une augmentation de l'immersion est associée à une diminution du nombre de fixations oculaires au cours du temps. Or, dans notre expérience, la figure 123 (page 260) nous montre que l'évolution du nombre de fixations pour la mauvaise présentation se dégrade plus vite au cours du temps par comparaison avec la bonne présentation. De même, on constate que le nombre total de fixation est plus important pour les conditions immersives, c'est à dire pour la bonne

présentation et la bonne interface que pour les plus mauvaises. Deux hypothèses peuvent expliquer une inversion de ces résultats par rapport à l'étude de Jennett et al. (2008). La première, vient de l'ambiguïté de l'interprétation cognitive d'une durée longue de fixation (ou d'un plus petit nombre de fixations totales, les deux étant liées).Cela peut être ,soit un indicateur d'un intérêt notable, soit un indicateur d'une situation complexe, qui nécessite en outre plus de temps pour être traité cognitivement (Jacob & Karn, 2003; Just & Carpenter, 1976). Ainsi, une plus grande durée de fixation pour les mauvaises conditions peut être due à un effort de concentration plus grand, ou, au contraire, une plus grande apathie. De même, la dégradation plus rapide du nombre de fixations au cours du temps pourrait être causée par un effort plus grand et nécessaire pour comprendre la présentation dans des conditions dégradées. On peut supposer ainsi que cette concentration attentionnelle forcée explique la corrélation significative observée entre des fixations plus longue et un plaisir accru plus faible. L'autre hypothèse découle directement du format visuel présenté aux utilisateurs. On peut supposer que la bonne présentation et la bonne interface offre plus de stimulations visuelles, par un cadre vidéo plus grand, un montage plus dynamique et de nombreuses ressources supplémentaires à disposition, ce qui a pour effet, d'aiguiser la curiosité et de préserver le tonus oculaire. Ces deux hypothèses s'appuient sur les deux sources cognitives de l'attention : « *top-down* », pour celles provenant de la motivation interne du sujet, et « *bottom-up* », pour celle provenant du stimulus en lui-même (Pashler, Johnstone, & Ruthruff, 2001). La difficulté ici est d'interpréter ces indicateurs en lien avec l'immersion, dans le sens où ils ne varient pas seulement dans un seul sens en fonction du niveau d'intérêt (une fixation longue peut indiquer une difficulté cognitive déplaisante autant que l'intérêt), mais également en fonction de la structure visuelle de l'expérience d'interaction.

Cette étude nous offre également un grand nombre d'informations sur la pertinence des indicateurs composites et souligne l'intérêt de combiner des indicateurs de modalités différentes en fonction de leurs fiabilités respectives. Dans l'expérience 1, deux indicateurs composites ont été construit. Les analyses de leur qualité psychométrique, à partir d'un modèle multi-facettes pour 15 utilisateurs, nous montrent que l'indicateur agrégé de la modalité textuelle ($\Phi = 0{,}859$) et vidéo ($\Phi = 0{,}807$) présente une fiabilité remarquable ($\Phi = 0{,}956$). Cela démontre la pertinence de combiner des indicateurs de fiabilité proches mais de modalités très différentes, car cela permet aux données de se compléter par des informations de sources distinctes. On constate néanmoins que l'indicateur composite général, qui agrège les trois modalités (textuelle, vidéo et subjectif), présente des performances qui se confondent presque totalement avec l'indicateur composite objectif ($\Phi = 0{,}956$). Ainsi, on peut présumer que l'indicateur subjectif n'a pas ajouté d'informations complémentaires par rapport aux deux autres modalités, dont la facette aurait déjà réduit drastiquement la marge d'erreurs. Dans l'expérience 2, aucun indicateur composite n'a été construit car, et cela constitue la limite principale de l'approche multimodale, il n'y avait pas d'indicateurs qui était, à la fois, de modalités significativement différentes et possédant des niveaux de fiabilité assez proche pour que leur combinaison soit productive.

Enfin, cette étude nous a permis également d'avoir un retour critique sur la pertinence de la procédure de validation multi-facettes dans le domaine des IHMs et particulièrement pour estimer la pertinence des mesures de l'immersion. Ainsi, cette approche s'est montrée

particulièrement efficace pour comparer les indicateurs entre eux et leurs combinaisons. Il s'agit donc d'une méthode permettant de démêler certaines informations contradictoires dans la littérature, en disposant de critères quantitatifs sur leurs fiabilités respectives. De plus, la facette utilisateur, considérée comme particulièrement importante dans le cadre de la mesure de l'expérience utilisateur, a également pu être examinée précisément grâce à ce type d'étude. Le taux de variance d'erreur pour cette facette a été de 21,7% pour l'expérience 1 et de 24,4% pour l'expérience 2, ce qui en fait un levier intéressant pour améliorer la qualité de la mesure. Cela montre également que la disparité d'expérience est importante d'un participant à l'autre. Cela est en raccord avec notre hypothèse formulée précédemment, à savoir que, la mesure de l'expérience utilisateur nécessite un nombre important de participants, car faisant appel à des dimensions plus subjectives et personnelles de l'interaction. Enfin, l'utilisation d'une procédure de validation multi-facettes sur les données des expériences 1 et 2 nous a montré tout l'enjeu méthodologique autour des procédures de mesure. En effet, dans la première expérience, le protocole testé se fondait sur la discrimination d'éléments selon un plan d'expérience croisé, c'est à dire un test où tous les utilisateurs devaient évaluer toutes les vidéos. Dans la deuxième expérience, le protocole testé se fondait sur la discrimination d'éléments selon un plan d'expériences croisées et emboitées, c'est à dire un test où tous les utilisateurs devaient évaluer toutes les vidéos mais seulement une des trois interfaces à la fois. Cette différence a montré un gros impact sur la fiabilité du protocole de test, bien moins précis dans le second cas. En effet, il est plus facile de mettre au point un protocole de test fiable dont le but est de tester l'immersion de plusieurs vidéoconférences visionnées par tous (condition croisée), qu'un protocole de test où cette exercice de discrimination prend en compte l'interface de visionnage, qui testée seulement une partie de l'échantillon utilisateur (condition emboitée)

CONCLUSION

Nous avons vu que de nombreux bouleversements du domaine ont entrainé l'émergence d'un nouveau paradigme d'évaluation, basé sur l'expérience utilisateur. Ce courant se différencie fondamentalement de ses prédécesseurs par son caractère holistique, subjectif et positif. Un grand nombre de méthodes de mesures ont ainsi été mises au point afin de mesurer cette expérience utilisateur, mais demeurent immatures, voire contradictoires. De plus, les caractéristiques même de l'expérience utilisateur conduisent à une perte de fiabilité lors de l'exercice de sa mesure. C'est pourquoi nous avons entrepris dans cette thèse un travail d'investigation des pistes existantes visant à la création de mesures robustes de l'expérience utilisateur, en se basant sur le cas d'usage de l'immersion.

Pour cela, nous nous sommes basés sur deux approches méthodologiques prometteuses : l'approche multimodale et l'approche multifacette. Au travers de trois études, cette posture épistémologique nous a permis de dégager de nombreux résultats et d'avoir un regard critique sur ces approches. Une synthèse de ces résultats est présentée ici autour de trois thèmes :

- **La mesure de l'immersion** : qui présente les principaux enseignements de la thèse obtenue pour les mesures testées et créées
- **L'approche multimodale** : qui présente les principaux enseignements de la thèse obtenus au contact de cette méthode et la pertinence des mesures composites créées
- **L'approche multifacette** : qui présente les principaux enseignements de la thèse obtenus par l'utilisation de la théorie de la généralisabilité

De plus, chacune des parties présentera les limites de l'approche développée et les perspectives existantes pour les surmonter. Enfin, nous finirons par quelques dernières considérations générales en guise de conclusion.

La mesure de l'immersion

L'immersion fait partie de ces construits qui rassemble de plus en plus de chercheurs dans le domaine de l'expérience utilisateur. Cette quête s'apparenterait presque à la recherche d'un nouveau facteur « *g* », propre à l'expérience utilisateur et permettant d'être en phase à tout moment avec le niveau de qualité subjective au cœur d'une expérience interactive. Ce construit, synthétique par nature, est particulièrement intéressant pour la mise au point d'une mesure globale de l'UX. En effet, résultant de l'impact d'une série de qualités UX en amont, ils sont moins soumis à la contingence.

La posture de la thèse a été de cerner l'immersion à partir d'une conceptualisation et d'une opérationnalisation solide. De plus, dans le contexte de la mesure d'un construit abstrait comme l'UX, nous pensons qu'il convient d'ajouter une dernière étape, celle de la triangulation, en s'appuyant sur un large spectre d'indicateurs, d'ordre physiologiques, comportementaux et

auto-rapportés. C'est pourquoi ces indicateurs ont été soigneusement évalués dans leurs capacités à cerner l'état d'immersion de l'utilisateur.

La première approche testée, la plus commune, a été d'utiliser une échelle auto-rapportée sous la forme d'un questionnaire. Nous nous sommes appuyés pour cela sur les dimensions de l'immersion les plus consensuelles et synthétiques afin d'obtenir une mesure la plus stable et la moins soumise à la contingence. De plus, son format court (9-items) ne rend pas la tâche d'auto-évaluation trop contraignante et peut donc être utilisé dans une variété de contextes. Les résultats ont montré de bonnes performances et confirment la validité de cette approche. Pour aller plus loin, nous avons également testé une échelle avec 1 item unique, afin de disposer d'un outil pouvant s'embarquer dans une application et viser une mesure au plus proche de l'interaction, voire en temps réel. Les résultats obtenus dans le cadre de cette thèse ont été très encourageant, en présentant des qualités psychométriques supérieures à toutes les autres mesures testées, notamment du point de vu de la sensibilité. Ainsi, ce type de mesures peut également ouvrir des pistes de recherche pour étudier la dimension temporelle et dynamique de l'expérience utilisateur, plébiscité par de nombreux auteurs (Hassenzahl & Sandweg, 2004; Nacke & Lindley, 2008; Ravaja, 2008) Néanmoins, il convient de s'assurer que l'utilisateur comprenne bien préalablement ce qu'il doit auto-rapporter. En effet, il est possible que la bonne performance obtenue dans cette recherche ait bénéficié par effet de halo du remplissage du questionnaire précédent. Dans tous les cas, ce type de mesure, reste incontournable, ne serait-ce que pour appuyer le développement d'autres mesures, étant donné ses bonnes qualités psychométriques et son cout réduit de mise en œuvre.

Une autre approche, dans la continuité de l'analyse des informations communiquées par l'utilisateur, est la mesure de l'immersion par les traces écrites. Cette méthode a montré de bons résultats pour la mesure de l'immersion, indirectement par le jugement de la valence affective des traces. Néanmoins, si cette inférence a été permise grâce au jugement d'un panel d'experts, le chemin est encore long avant de pouvoir le réaliser automatiquement par des algorithmes auto-apprenants. Ainsi, les utilisateurs de ces techniques devront se baser, comme actuellement, sur un nombre très important d'utilisateurs pour être capable de détecter des informations intéressantes, dans un flot de données non contrôlées et difficilement interprétables précisément par la voie algorithmique (Huron et al., 2013; Kramer, 2010).

L'approche qui nous a semblé être une des plus prometteuses à investiguer a été la mesure de l'immersion par les indices non verbaux, tels que les expressions faciales et corporelles. En effet, les résultats obtenus dans cette thèse ont montré que les enregistrements de l'utilisateur, grâce à la webcam, ont livré une grande quantité d'indice sur l'état d'immersion de l'utilisateur, ce qui a permis de le mesurer avec une bonne précision. Ainsi, des études plus précises sur ces indices permettraient de nourrir les algorithmes de reconnaissance visuels, qui sont pour le moment plutôt tournés sur l'analyse des émotions primaires.

D'un autre côté, les mesures physiologiques utilisées dans la thèse ont mis à jour de nombreuses difficultés qui conviendra de surmonter avant de pouvoir exploiter ces mesures dans le cadre de l'évaluation de l'immersion. En effet, ces mesures ont manqué principalement de diagnosticité (habilité à cibler précisément un état psychologique sans être affecté par d'autres) et de fiabilité (consistance d'une inférence psycho-physiologique entre les différents individus

et environnements). Ces qualités sont pourtant indispensables pour permettre une mesure valide d'un état psychologique donné (Fairclough, 2009). En effet, la conductance de la peau a manqué de diagnosticité dans l'expérience 1 de l'étude 3, dans le sens où il n'y a pas été possible de distinguer les réponses électrodermales provenant des mouvements physiques (clavier) de celles issues d'une activé mentale associée à l'expérience de visionnage de la vidéoconférence. De même, la conductance de la peau n'a pas permis d'associer les niveaux de sur/sous stimulation avec un niveau d'immersion correspondant. De plus, la conductance de la peau a manqué de fiabilité, dans le sens où les différences interindividuelles naturelles étaient encore importantes, malgré de nombreuses tentatives de normalisation du signal. De l'autre côté, la mesure du rythme cardiaque n'a donné aucun résultat, et il n'y a pas été possible d'en expliquer précisément les causes (manque de diagnosticité ? fiabilité ? sensibilité ?). Ainsi, nous pensons qu'il est encore trop top pour utiliser ces indicateurs dans le cadre d'une mesure de l'immersion. Deux pistes de recherche peuvent être toutefois envisagées pour résoudre certains des obstacles rencontrés. La première est le croissement des signaux physiologiques entre eux pour augmenter leurs niveaux de diagnosticité. Dans cette approche, Picard, Vyza et Healey (2001) ont réussi la reconnaissance de huit émotions avec une précision de 81%, en combinant les mesures de la respiration, de la pression sanguine, de la conductance de la peau et des expressions faciales. Néanmoins, celle suppose de multiplier les capteurs physiologiques, ce qui a un coût et peut rendre la situation plus invasive encore pour l'utilisateur. En parallèle donc, un travail est à réaliser sur la réduction de l'invasivité de tels systèmes, tout en en améliorant leurs fiabilités et résistances aux artéfacts de mesures (tels que les mouvements). L'autre piste de recherche pour améliorer la fiabilité est la mise au point de systèmes de mesure normalisés par rapports aux caractéristiques physiologique de chaque utilisateur. En effet, nous pourrions imaginer un outil dont l'algorithme s'affinerait au cours du temps, en apprenant des variations physiologiques individuelles, comme le font les logiciels de reconnaissance vocale actuels. De même, des procédures puissantes de calibration pourraient y être associées, pour gommer plus efficacement encore les particularités physiologiques individuelles.

L'approche multimodale

Dans le cadre de cette thèse, nous avons mis à l'épreuve l'approche multimodale dans le cadre de l'amélioration de la mesure de l'immersion. Pour le moment, cette approche est peu utilisée pour plusieurs raisons : elle est plus couteuse à mettre en place et nécessite de maitriser toutes les techniques de recueil, d'analyse et de croisement des données associées. Ainsi, la majorité des praticiens préfèrent utiliser une seule mesure fiable, tel que le questionnaire d'immersion, bien que présentant de nombreux biais inhérents à la méthode (l'effort d'extraction de l'utilisateur de l'expérience, la déformation de l'appréciation à cause de la remémoration d'un état subjectif passé, …). Néanmoins, la multiplication des dispositifs de captures intégrées, couplée à l'abondance de plus en plus importante des traces utilisateurs, fait que cette approche a tendance de plus en plus à se démocratiser. En effet, nous voyons déjà émerger avec le *Big Data* une automatisation de l'évaluation quantitative, bien que souvent couplée à la création d'algorithmes, dont la logique d'action est malheureusement embarquée à l'intérieur de nombreuses « *boites noires* ».

En résumé, nous pensons que combiner des informations de sources diverses nous permet d'appréhender une entité de manière plus fiable et complète. En effet, l'indépendance des sources permet de neutraliser certains biais inhérents à l'acquisition de certaines données et d'obtenir une mesure globale plus fiable (Creswell, 2003; Wilson, 2006). Les résultats de cette thèse tendent à confirmer cette affirmation. En effet, nous avons vu, dans l'étude 2, l'intérêt de combiner divers composants objectifs de la mesure de l'utilisabilité, telle que la réussite à une tâche (composant d'efficacité) et la performance à une tâche (composant d'efficience). Cette fusion a ainsi augmenté la finesse de la mesure, en gommant le biais propre à la différence de persévérance des utilisateurs. Le même résultat a été à l'œuvre avec la fusion des indicateurs objectifs et subjectifs. D'une fiabilité quasi-identique, leur fusion nous a offert un composé d'une précision significativement plus grande. De même, dans l'expérience 1 de l'étude 3, les analyses psychométriques de l'indicateur composite, agrégeant la modalité textuelle et vidéo, ont présenté une fiabilité remarquable pour la mesure de l'immersion, en comparaison de chacun des indicateurs séparés. Ainsi, de manière plus générale, nous constatons que la combinaison d'indicateurs pointant sur un construit commun, peu corrélés entre eux, mais de fiabilité proche, augmente la fiabilité de l'indicateur ainsi combiné.

On constate néanmoins que l'indicateur composite général, qui agrège les trois modalités (textuelle, vidéo et subjectif), a présenté des performances qui se confondent presque totalement avec l'indicateur composite objectif. Ainsi, on peut présumer que l'indicateur subjectif n'a pas ajouté d'informations complémentaires par rapport aux deux autres modalités. De même, dans l'expérience 2, aucun indicateur composite n'a été construit, car n'y avait pas d'indicateurs de modalités différentes possédant des niveaux de fiabilité assez proches. Ainsi, deux leçons peuvent être tirées de ces résultats. La première, c'est que la modalité des indicateurs à combiner doit être suffisamment éloignée l'une de l'autre pour que leur fusion tire parti de leurs complémentarités et du contrebalancement de leurs biais respectifs. La deuxième, c'est que les indicateurs ne doivent pas être de fiabilités significativement différentes, pour ne pas annuler simplement le bénéfice de leurs combinaisons. Les recherches multimodales futures pourraient donc être menées pour maximiser le potentiel de ces deux facteurs.

Enfin, un travail pourrait être réalisé sur le choix de la procédure de pondération des indicateurs composites. Dans cette thèse, nous avons fait le choix de pondérer de manière égale tous les éléments des indicateurs composites. Or, d'autres procédures, plus fines, pourraient être mises en œuvre, comme celle proposée sur la théorie de la généralisabilité multivariée (Webb & Shavelson, 1981) et qui propose de pondérer les différents éléments de l'indicateur composite en fonction de leurs coefficients de généralisabilité respectifs. Enfin, ce processus peut être complété par des analyses d'incertitude et de sensibilité pour tester la robustesse des indicateurs composites construits (Cherchye et al., 2008; Saisana, Saltelli, & Tarantola, 2005).

L'approche multi-facettes

Enfin, dans le cadre de cette thèse, nous avons testé une nouvelle approche de validation psychométrique : l'approche multi-facettes. En effet, le domaine actuel se cantonne encore majoritairement aux outils de validation psychométrique bi-facettes issus de l'héritage historique plutôt que d'un véritable choix épistémologique. Or, nous pensons que l'application

de l'approche multi-facettes dans le domaine de l'évaluation des systèmes numériques est plus pertinente, dans le sens où la mise au point d'un protocole d'évaluation dans ce domaine vise à discriminer les applications entre elles, et non les individus qui les utilisent. En effet, par rapport à la théorie classique de la mesure, la théorie de la généralisabilité permet de s'intéresser à la mesure d'entités autres que des individus, tels que des items, des objets ou des méthodes. De plus, cette méthode nous permet d'étudier précisément la facette utilisateur (cachée dans les protocoles psychométriques classiques), qui, à cause de la subjectivité lors d'une interaction, pourrait contribuer grandement à la variabilité de l'évaluation, d'autant plus avec l'avènement de l'UX.

L'utilisation de cette nouvelle approche de cette thèse a permis ainsi de démontrer sa pertinence en captant l'influence des paramètres principaux agissant sur la fiabilité de la mesure et couramment occultée dans les recherches méthodologiques actuelles. Nous avons ainsi pu constater l'impact de certains facteurs comme le nombre de tâches (Etude 2) et d'utilisateurs (Etude 1, 2, 3). Nous avons également vu qu'il est possible d'estimer précisément le nombre d'éléments nécessaires dans chaque facette pour que le test soit considéré comme suffisamment fiable, en se basant sur leurs instabilités respectives. Ainsi, nous avons observé dans l'étude 2 que les performances et l'utilisabilité perçues varient en quantités importantes en fonction de la tâche, bien plus encore que d'un utilisateur à l'autre. De même, cela nous a permis de comparer la précision des différents indicateurs de l'immersion testée, ainsi de leurs dérivés composites dans l'étude 3. Tous ces éléments nous ont permis d'estimer avec finesse la qualité des protocoles de mesure, en possédant un modèle psychométrique véritablement adapté à la structure multi-facettes de l'évaluation dans le milieu des IHM.

Enfin, il convient d'aborder certaines limites et contraintes de cette approche. Les premières sont d'ordre statistique. En effet, la réalisation d'une étude de généralisabilité ne permet pas d'inclure des données dont certaines sont manquantes (par exemple les données d'un utilisateur dont ils manquent seulement des indicateurs cardiaques). Cela nous a conduit à écarter de nombreuses données utilisateur de l'étude, car, au vu du grand nombre de données récoltées, certaines d'entre elles n'ont pas été correctement recueillies. De même, la théorie de la généralisabilité ne permet pas de tester des plans non équilibrés (ex : tester une facette avec une mesure à *x* items et une mesure à y items), ce qui nous a forcé, soit à construire plusieurs études G (Etude 2), soit à agréger certains niveaux (Facette juge de l'expérience 1 de l'étude 3). Cela est problématique car cela nous empêche d'avoir une vue globale sur le protocole de mesure avec la totalité des forces de variabilités à l'œuvre. Néanmoins, des travaux sont actuellement à l'œuvre pour permettre à la théorie de la généralisabilité de surmonter ces deux limites (Brennan, 2013). La deuxième contrainte a été d'ordre méthodologique. En effet, une des contraintes du modèle concerne la collecte des éléments des facettes de différenciation. Théoriquement, les éléments de cette facette doivent être tirés au hasard et suffisamment nombreux pour représenter la variabilité naturelle existante. Pour l'étude 2, cela n'a pas posé de problème pour la facette « site universitaire », car une liste complète représentant tous les sites a pu être trouvée, et dont les 5 éléments de l'étude ont pu y être tirés au sort. Néanmoins, cela n'a pas pu être le cas pour l'étude 3 car les conditions immersives ont dû être crées de toute pièce, n'existant pas de bibliothèque partagée et validée à ce jour. Le risque encouru est que les coefficients en valeur absolue ne soient pas représentatifs, car reposant sur une facette de

différentiation dont la variance est, soit plus importante, soit moins importante, que ce qu'elle devrait être. De manière plus générale, le domaine de l'immersion aurait à gagner à disposer d'un ensemble de situations, plus ou moins immersives, pouvant servir de situation témoin entre les différentes études. Cela permettrait, comme cela se fait dans d'autres domaines, de disposer d'étalons pour tester différentes procédures de test et de les comparer entre elles.

Dernières considérations générales

Toute approche seule, reposant sur un outil, une méthode, ou une théorie, possède inévitablement des points d'ombres. C'est pour cela que nous nous sommes inscrits dans le courant du « *critical multiplism* ». Ce mouvement milite pour l'utilisation de plusieurs approches pour répondre à une question particulière, mais pas dans le sens d'un activisme aveugle du type « *tout convient* », mais au contraire dans un sens critique et systématique. C'est-à-dire que ce n'est pas le nombre d'options mis en œuvre qui importe mais plutôt « *que les différentes options sélectionnées possèdent des biais qui opèrent dans des directions différentes* » (Shadish, 1986). Ainsi, cette approche atteint son but quand la diversité injectée à la résolution du problème particulier souligne sa complexité et révèle les biais de chacune des techniques isolés.

Nous pensons également que cette approche évite tout aveuglement et tend à faire dialoguer les savoirs et disciplines entre elles. En effet, nous ne croyons pas que le champ de l'ergonomie puisse se passer à ce jour des nouvelles approches informatiques et algorithmes pour avancer, de même que l'informatique puisse se passer du reste du monde académique par l'utilisation d'algorithme surpuissant, auto-apprenant et a-théorique… ou tout du moins pas encore avant quelques décennies. Nous pensons donc que l'avènement de la profusion des traces utilisateurs, leur recueil et analyse devra se faire conjointement entre informaticiens, ingénieurs et psychométriciens.

Enfin, nous pensons également que le développement de cette approche ne pourra se faire sans le consentement des utilisateurs eux-mêmes. En effet, le risque est de voir la captation de ces données uniquement dans la main de quelques acteurs, sans aucune transparence et partage des connaissances. Cela rejoint la critique de Boullier sur la position de Colin & Verdrier (2012) dans leur livre, l'« *Age de la multitude* » où ces derniers s'exclamait que : « *les entreprises doivent s'exercer à capter la puissance de cette multitude » (p.82*), seule source de créativité, qu'il convient non pas de valoriser comme bien commun, mais de capter au profit de certains acteurs[94]. L'ouverture de ces données est donc un enjeu important mais tout autant que la transparence sur leurs utilisations et le respect de l'éthique. Enfin, nous pensons que l'exploitation de ces données doit bénéficier en premier aux utilisateurs, en leur offrant de meilleures expériences d'interaction et/ou une connaissance approfondie d'eux-mêmes (cf. mouvement du *quanfied self*). Nous finirons donc par ces mots de Boulier résumant ces deux

[94] Tirée de la réponse de Boullier (2012)
http://www.internetactu.net/2012/09/07/l%E2%80%99age-de-la-predation/

forces en lutte : « *le conflit est là, ouvert, violent, entre les tenants de la captation de la créativité collective (de Apple à Amazon et à « toutes les plates-formes ») et les réseaux de la culture libre qui bataillent chaque jour pour préserver la neutralité du net, l'ouverture des données, le refus de la fermeture du code et les contrôles étatiques et financiers sur la circulation des idées.* »

ANNEXES

ANNEXE 1 - PRE-ETUDE JEUX VIDEO

Introduction

En étudiant la facette « utilisateur » de l'évaluation, des travaux dans le domaine des jeux vidéo ont montré une fiabilité inter-juge faible lors d'évaluations à partir de listes d'heuristiques (Korhonen, Paavilainen, & Saarenpää, 2009; White, Mirza-babaei, McAllister, & Good, 2011)**.** Cet « *effet évaluateur* » a été également mis à jour récemment dans les méthodes d'inspection plus traditionnelles (Hertzum & Jacobsen, 2003). Ces derniers listent trois causes d'inconstance inter-évaluateurs dans l'inspection experte : (1) des buts d'analyses vagues, (2) des procédures d'analyses vagues, (3) des critères d'inspection vagues. Cette troisième cause est liée directement à la construction des méthodes heuristiques (Cockton et al. 2013). Pour White et al. (2010), la faible fiabilité inter-juge lors de l'évaluation experte de jeux-vidéo est causée par la complexité de l'investigation associée à certaines heuristiques. En effet, en se basant sur la taxonomie de Cockton and Woolrych (2001), White et al. stipulent que les modèles d'inspection dans les jeux vidéo utilisent majoritairement des heuristiques *« à construire* » (« *constructable* heuristics »), qui nécessitent une longue interaction avec le jeu et d'être contrôlés sur de multiples objets, ce qui explique la faible consistance inter-juge observée. Comme la plupart des domaines associés aux IHM, les modèles d'inspection dans les jeux vidéo sont passés d'une évaluation centrée sur l'utilisabilité à l'évaluation prenant en compte toute l'expérience utilisateur. C'est pourquoi il nous a semblé intéressant d'étudier l'impact du type d'heuristique sur la variabilité inter-juges.

Hypothèse de travail

Nous avons utilisé le cas d'usage de l'évaluation des jeux vidéo pour tester formellement l'hypothèse selon laquelle **l'utilisation de critères d'évaluation centrés sur l'expérience utilisateur (plus subjectif) entraîne plus de disparité de jugement entre les évaluateurs que des critères centrés sur les éléments du produit (plus objectif).**

Méthode et procédure

Pour tester cette hypothèse, nous avons basé notre analyse sur des données issues d'une enquête réalisée sur 120 joueurs dont certains des résultats ont été par ailleurs publiés (Rodio & Bastien, 2013).

L'analyse a été affinée grâce à deux catégories de joueurs et de trois types de jeu différents. Les joueurs interrogés ont été catégorisés en « *amateurs* » et « *e-sportifs* ». Les joueurs « *amateurs* » sont des joueurs qui jouent pour leur propre plaisir sans se prendre trop au sérieux alors que

les « *e-sportifs* » s'épanouissent en compétition et font tout leur possible pour améliorer leur performance. Afin de représenter les genres vidéo-ludiques, trois genres très répendus ont été retenus : les RTS (« *Real Time Strategy* »), les MMORPG (« *Massively Multi-player Online Role-Playing Game* ») et les FPS (« *First Person Shooter* »). Pour représenter ces trois genres, trois jeux ont été choisis : « *C&C : Alerte Rouge 3* », « *Call of Duty 4: Modern Warfare* » et « *World of Warcraft* ». Ces trois jeux représentent fidèlement leur genre respectif et possèdent toutes les caractéristiques nécessaires pour répondre aux objectifs de cette étude (leurs communautés de joueurs est assez grande pour faciliter leur accès, leurs populations est assez semblable pour pouvoir être comparées fidèlement, et, pour chacune, assez diversifiées pour pouvoir interroger tous types de joueurs). Enfin, un modèle d'inspection a été choisi : la nouvelle version du modèle PLAY de Desurvire et Wiberg (2009). Composé de 47 critères ergonomiques, ce modèle se divise en trois grandes familles : « *Game Play* », « *Usability & Game Mechanics* » et « *Coolness/Entertainment/Humor/Emotional Immersion* ». L'expérience a consisté pour chaque joueur (Amateur ou e-Sportif), à évaluer un jeu donné (« *C&C : Alerte Rouge 3* », « *Call of Duty 4: Modern Warfare* » et « *World of Warcraft* ».) à l'aide des 47 critères d'évaluation.

Pour l'analyse statistique réalisée dans le cadre de cette thèse, seules les catégories « Game Play » (20 items) et « *Usability & Game Mechanics* » (21 items) ont été conservées. Les items de la catégorie « Game Play » représente l'expérience du joueur, tel que le niveau de challenge, le rythme de jeu ou encore la perception de contrôle : ce sont des critères proprement UX. Les items de la catégorie « *Usability & Game Mechanics* » représentent les éléments du jeu et de l'interface, tels que la mise en page de l'écran, la navigation ou encore les mécanismes de prévention des erreurs : ce sont des critères centrés sur les éléments de jeu. Pour chacun de ces 41 items, 6 scores de variance ont été calculés :

- L'écart type des items pour l'évaluation du jeu par les 40 joueurs de « *C&C : Alerte Rouge 3* »
- L'écart type des items pour l'évaluation du jeu par les 40 joueurs de « *Call of Duty 4: Modern Warfare* »
- L'écart type des items pour l'évaluation du jeu par les 40 joueurs de « *World of Warcraft* »
- L'écart type total des items par l'évaluation des 60 joueurs amateurs (moyennes des écarts types pour les 3 jeux)
- L'écart type total des items par l'évaluation des 60 joueurs eSportifs (moyennes des écarts types pour les 3 jeux)
- L'écart type total des items (moyenne des écarts types pour les 3 jeux)

Puis, 6 tests de Student ont été réalisés sur ces données pour tester la significativité des différences en termes de variance pour les items de type « *UX* » ou « *Eléments de jeu* »

Résultats et conclusion

	Type d'item	N	Moyenne	σ	t	Sig.
Ecart type des items pour l'évaluation de « C&C : Alerte Rouge 3 » (RTS)	Eléments de jeu	21	0,87	0,174	-2,983	.005
	UX	22	1,05	0,213		
Ecart type des items pour l'évaluation de « Call of Duty 4: Modern Warfare » (FPS)	Eléments de jeu	21	0,83	0,150	-5,378	,000
	UX	22	1,07	0,145		
Ecart type des items pour l'évaluation de « Call of Duty 4: Modern Warfare » (MMORPG)	Eléments de jeu	21	0,85	0,175	-3,512	,001
	UX	22	1,04	0,178		
Ecart type des items pour les joueurs amateurs	Eléments de jeu	21	0,85	0,158	-3,726	,001
	UX	22	1,01	0,113		
Ecart type des items pour les joueurs eSportifs	Eléments de jeu	21	0,82	0,131	-5,552	,000
	UX	22	1,07	0,157		
Moyenne des écarts type pour les 3 jeux	**Eléments de jeu**	**21**	**0,85**	**0,125**	**-2,621**	**,012**
	UX	**22**	**1,00**	**0,245**		

La table ci-dessus nous montre que la consistance inter-évaluateur des critères d'évaluation centrée sur l'expérience du joueur (UX) est plus faible que celle des critères d'évaluation centrés sur les éléments de jeu ($n = 43, p = .012$), et cela, quel que soit le profil du joueur interrogé ($p < .01$, pour les joueurs amateurs et pour les joueur e-sportifs) ou le type de jeu ($p < .01$ pour les jeux de type RTS, FPS ou MMORPG).

Ces résultats sont très intéressants car le domaine des jeux vidéo a été un des premiers à se tourner vers la mesure de l'UX. Cela nous confirme l'importance dans ce domaine de disposer de données sur la facette « utilisateur » tant cette tâche demande une estimation importante de la part subjective du jugement. Nous pouvons conclure plus généralement que l'utilisation de critères d'évaluation centrés sur l'expérience utilisateur (plus subjectif) entraîne plus de disparité de jugement entre les évaluateurs que des critères centrés sur les éléments du produit (plus objectif).

ANNEXE 2 – MATERIEL DE L'ETUDE 2

A. Email de participation à l'étude

Objet : invitation à un test utilisateur sur des sites web universitaires

Madame, Monsieur,

L'équipe ETIC de l'Université de Lorraine, en partenariat avec Alcatel-Lucent, ont mandaté la société Insydelabs (spin-off d'ETIC) pour vous inviter à réaliser un test utilisateur sur plusieurs sites web universitaires. A partir d'un outil en ligne, vous aurez à effectuer une série de tâches sur différents sites web. Vous pourrez terminer chaque tâche en cliquant sur le bouton « j'ai réussi » ou le bouton « j'ai échoué » qui se trouvera en haut de votre navigateur. Veuillez noter que nous n'évaluerons pas votre performance aux tâches, l'unique objectif étant d'améliorer les sites web en questions.

Pendant ce test, votre écran et vos actions sur le navigateur seront enregistrés. Veuillez également choisir un endroit où vous ne serez pas dérangé ou interrompu durant le test. La durée de ce test est d'une heure environ (Questionnaire inclus).

Afin de participer, merci d'ouvrir le lien ci-dessous avec le navigateur Firefox : http://app.insydelabs.com/evadis/testeur/invitation?token=9a23b193a389ebc35f7b085fbce6e6149c853f20

Enfin, une fois le test réalisé, nous vous demanderont de remplir le questionnaire disponible à cet adresse :

https://docs.google.com/spreadsheet/viewform?formkey=dG42VUxtWG9rbVliSERsNDQweXYxYVE6MQ#gid=0.

Merci par avance pour votre participation,

Sincères salutations,

Florentin Rodio

Doctorant ETIC/Alcatel Lucent

Université de Lorraine

B. Questionnaire

- Adresse email (Champ texte)
- Age (Champ numérique)
- Sexe (bouton radio)
    - Homme
    - Femme
- Année d'étude universitaire (Bouton radio)
    - Licence première année
    - Licence deuxième année
    - Licence troisième année
    - Master première année
    - Master deuxième année
- A quel fréquence vous connectez vous sur le site de votre université ?
    - Tous les jours
    - Au moins une fois par semaine
    - Au moins une fois par mois
    - Quelques fois dans l'année
    - Jamais
- Vous êtes vous déjà connecté sur le site de l'université de université d'Aix-Marseille?
    - Oui, plusieurs fois
    - Oui, une fois seulement
    - Non, jamais
    - Je ne me rappelle plus
- Vous êtes vous déjà connecté sur le site de l'université de université de Caen?
    - Oui, plusieurs fois
    - Oui, une fois seulement
    - Non, jamais
    - Je ne me rappelle plus
- Vous êtes vous déjà connecté sur le site de l'université de université du Sud Toulon-Var?
    - Oui, plusieurs fois
    - Oui, une fois seulement
    - Non, jamais
    - Je ne me rappelle plus
- Vous êtes vous déjà connecté sur le site de l'université de université de Brest?
    - Oui, plusieurs fois
    - Oui, une fois seulement
    - Non, jamais
    - Je ne me rappelle plus
- Vous êtes vous déjà connecté sur le site de l'université de université de Toulouse Capitole?
    - Oui, plusieurs fois
    - Oui, une fois seulement
    - Non, jamais
    - Je ne me rappelle plus

C. Questionnaires SUS

Questionnaires de satisfaction - Sites web universitaires

Pour terminer cette expérience, veuillez remplir les questionnaires suivant pour chacun des 5 sites universitaires précédemment explorés.

*Obligatoire

Veuillez indiquer votre adresse e-mail *

Il est important de renseigner la même adresse que lors du test précédent afin que nous puissions correctement lier les questionnaires suivants à celui-ci

Site web de l'université d'Aix-Marseille (http://www.univ-amu.fr/) *

Afin de mieux repondre à ce questionnaire, nous vous conseillons de compléter votre connaissance de ce site par une exploration libre de celui ci à l'adresse: http://www.univ-amu.fr/

	1- Pas du tout d'accord	2	3	4	5 - Tout à fait d'accord
1. Je pense que j'aimerais utiliser ce site fréquemment					
2. Je trouve ce site inutilement complexe					
3. Je pense que ce site est facile à utiliser					
4. Je pense que j'aurais besoin de l'aide d'un technicien pour être capable d'utiliser ce service					
5. J'ai trouvé que les différentes fonctions du site ont été bien intégrées					

	1- Pas du tout d'accord	2	3	4	5 - Tout à fait d'accord
6. Je pense qu'il y a trop d'inconstances dans ce site					
7. J'imagine que la plupart des gens serait capable d'apprendre à utiliser ce site très rapidement					
8. J'ai trouvé ce site très lourd à utiliser					
9. Je me suis senti très confiant lors de l'utilisation de ce site					
10. J'ai eu besoin d'apprendre beaucoup de choses avant de pouvoir utiliser correctement ce site					

Site web de l'université de Caen (http://www.unicaen.fr/) *

Afin de mieux repondre à ce questionnaire, nous vous conseillons de compléter votre connaissance de ce site par une exploration libre de celui ci à l'adresse: http://www.unicaen.fr/

	1- Pas du tout d'accord	2	3	4	5 - Tout à fait d'accord
1. Je pense que j'aimerais utiliser ce site fréquemment	○	○	○	○	○
2. Je trouve ce site inutilement complexe	○	○	○	○	○
3. Je pense que ce site est facile à utiliser	○	○	○	○	○
4. Je pense que j'aurais besoin de l'aide d'un technicien pour être capable d'utiliser ce service	○	○	○	○	○
5. J'ai trouvé que les différentes fonctions du site ont été bien intégrées	○	○	○	○	○
6. Je pense qu'il y a trop d'inconstances dans ce site	○	○	○	○	○
7. J'imagine que la plupart des gens serait capable d'apprendre à utiliser ce site très rapidement	○	○	○	○	○

	1- Pas du tout d'accord	2	3	4	5 - Tout à fait d'accord
8. J'ai trouvé ce site très lourd à utiliser					
9. Je me suis senti très confiant lors de l'utilisation de ce site					
10. J'ai eu besoin d'apprendre beaucoup de choses avant de pouvoir utiliser correctement ce site					

Site web de l'université du Sud Toulon-Var (http://www.univ-tln.fr/) *

Afin de mieux repondre à ce questionnaire, nous vous conseillons de compléter votre connaissance de ce site par une exploration libre de celui ci à l'adresse: http://www.univ-tln.fr/

	1- Pas du tout d'accord	2	3	4	5 - Tout à fait d'accord
1. Je pense que j'aimerais utiliser ce site fréquemment					
2. Je trouve ce site inutilement complexe					
3. Je pense que ce site est facile à utiliser					
4. Je pense que j'aurais besoin de l'aide d'un technicien pour être					

	1- Pas du tout d'accord	2	3	4	5 - Tout à fait d'accord
capable d'utiliser ce service					
5. J'ai trouvé que les différentes fonctions du site ont été bien intégrées					
6. Je pense qu'il y a trop d'inconstances dans ce site					
7. J'imagine que la plupart des gens serait capable d'apprendre à utiliser ce site très rapidement					
8. J'ai trouvé ce site très lourd à utiliser					
9. Je me suis senti très confiant lors de l'utilisation de ce site					
10. J'ai eu besoin d'apprendre beaucoup de choses avant de pouvoir utiliser correctement ce site					

Site web de l'université de Brest (http://www.univ-brest.fr/) *

Afin de mieux repondre à ce questionnaire, nous vous conseillons de compléter votre connaissance de ce site par une exploration libre de celui ci à l'adresse: http://www.univ-brest.fr/

	1- Pas du tout d'accord	2	3	4	5 - Tout à fait d'accord
1. Je pense que j'aimerais utiliser ce site fréquemment	○	○	○	○	○
2. Je trouve ce site inutilement complexe	○	○	○	○	○
3. Je pense que ce site est facile à utiliser	○	○	○	○	○
4. Je pense que j'aurais besoin de l'aide d'un technicien pour être capable d'utiliser ce service	○	○	○	○	○
5. J'ai trouvé que les différentes fonctions du site ont été bien intégrées	○	○	○	○	○
6. Je pense qu'il y a trop d'inconstances dans ce site	○	○	○	○	○
7. J'imagine que la plupart des gens serait capable d'apprendre à utiliser ce site très rapidement	○	○	○	○	○

	1- Pas du tout d'accord	2	3	4	5 - Tout à fait d'accord
8. J'ai trouvé ce site très lourd à utiliser					
9. Je me suis senti très confiant lors de l'utilisation de ce site					
10. J'ai eu besoin d'apprendre beaucoup de choses avant de pouvoir utiliser correctement ce site					

Site web de l'université de Toulouse Capitole (http://www.univ-tlse1.fr/) *

Afin de mieux repondre à ce questionnaire, nous vous conseillons de compléter votre connaissance de ce site par une exploration libre de celui ci à l'adresse: http://www.univ-tlse1.fr/

	1- Pas du tout d'accord	2	3	4	5 - Tout à fait d'accord
1. Je pense que j'aimerais utiliser ce site fréquemment					
2. Je trouve ce site inutilement complexe					
3. Je pense que ce site est facile à utiliser					
4. Je pense que j'aurais besoin de l'aide d'un technicien					

	1- Pas du tout d'accord	2	3	4	5 - Tout à fait d'accord
pour être capable d'utiliser ce service					
5. J'ai trouvé que les différentes fonctions du site ont été bien intégrées					
6. Je pense qu'il y a trop d'inconstances dans ce site					
7. J'imagine que la plupart des gens serait capable d'apprendre à utiliser ce site très rapidement					
8. J'ai trouvé ce site très lourd à utiliser					
9. Je me suis senti très confiant lors de l'utilisation de ce site					
10. J'ai eu besoin d'apprendre beaucoup de choses avant de pouvoir utiliser correctement ce site					

Envoyer

N'envoyez jamais de mots de passe via Google Forms.

D. Tâches et consignes personnalisées en fonction du site universitaire

Tâche 1 : Recherche de formation (1) sur le site d'Aix-Marseille (1)

- **Consigne :** Trouver le programme d'enseignements avec le volume horaire des cours pour le Master 1 en Energie nucléaire.
- **Adresse de départ :** http://www.univ-amu.fr/
- **Adresse de réponse**
  - (1) http://master-enam.univ-cezanne.fr/programme.html
  - (2) http://master-enam.univ-cezanne.fr/files/unites_enseignement.pdf sur la page http://master-enam.univ-cezanne.fr/

Tâche 2 : Recherche de sport (2) sur le site d'Aix-Marseille (1)

- **Consigne :** Trouver le(s) jour(s) de séance de Salsa proposé par le service universitaire d'activité sportive pour cette année.
- **Adresse de départ :** http://www.univ-amu.fr/
- **Adresse de réponse**
  - (1) http://siuaps.aix.univ-cezanne.fr/index_fichiers/Page390.htm

Tâche 3 : Recherche de bibliothèque (3) sur le site d'Aix-Marseille (1)

- **Consigne :** Trouver les horaires d'ouverture/fermeture de la bibliothèque de Théorie du Droit.
- **Adresse de départ :** http://www.univ-amu.fr/
- **Adresse de réponse**
  - (1) http://flora.univ-cezanne.fr/flora/servlet/LoginServlet onglet "informations pratiques">"Bibliothèques" (je n'arrive pas à récupérer l'url de ce sous-onglet)

Tâche 4 : Recherche de formation (1) sur le site de Caen (2)

- **Consigne :** Trouver le programme d'enseignements avec le volume horaire des cours pour le Master de Droit spécialisé en contentieux privé.
- **Adresse de départ :** http://www.unicaen.fr/
- **Adresse de réponse**
  - (1) http://www.unicaen.fr/master-pro-rech-droit-specialite-contentieux-prive-280762.kjsp?RH=1291198060074&ONGLET=3
  - (2) http://webetu.unicaen.fr/formations-et-etudes/formations-classees-par-type-de-diplome/master-pro-rech-droit-specialite-contentieux-prive-280762.kjsp?RH=1322040284275&ONGLET=3

Tâche 5 : Recherche de sport (2) sur le site de Caen (2)

- **Consigne :** Trouver le(s) jour(s) de séance d'escrime proposé par le service universitaire d'activité sportive pour cette année.
- **Adresse de départ :** http://www.unicaen.fr/
- **Adresse de réponse**
  - (1) http://webetu.unicaen.fr/medias/fichier/copie-de-maquette-planning-2012-2013_1349170665187-pdf?INLINE=FALSE sur la page: http://webetu.unicaen.fr/sport/activites-proposees/
  - (2) http://webetu.unicaen.fr/medias/fichier/2012-combat_1343981969761.pdf?INLINE=FALSE sur la page: http://webetu.unicaen.fr/sport/activites-proposees/
  - (3) http://webetu.unicaen.fr/medias/fichier/copie-de-maquette-planning-2012-2013_1349170629607-pdf?INLINE=FALSE sur la page: http://webetu.unicaen.fr/sport/

Tâche 6 : Recherche de bibliothèque (3) sur le site de Caen (2)

- **Consigne :** Trouver les horaires d'ouverture/fermeture de la bibliothèque Universitaire Santé
- **Adresse de départ :** http://www.unicaen.fr/
- **Adresse de réponse**
  - (1) http://scd.unicaen.fr/bibliotheques-/bibliotheques-universitaires-a-caen/bu-sante-174032.kjsp?RH=1254298250391

Tâche 7 : Recherche de formation (1) sur le site du Sud Toulon-Var (3)

- **Consigne :** Trouver le programme d'enseignements avec le volume horaire des cours pour le Master 1 en Information Communication.
- **Adresse de départ :** http://www.univ-tln.fr/
- **Adresse de réponse**
  - (1) http://formation.univ-tln.fr/Master-Information-Communication-1ere-annee.html#
  - (2) http://formation.univ-tln.fr/spip.php?action=telecharger&arg=135 sur la page http://formation.univ-tln.fr/Master-Information-Communication-1ere-annee.html#

Tâche 8 : Recherche de sport (2) sur le site du Sud Toulon-Var (3)

- **Consigne :** Trouver le(s) jour(s) de séance d'aviron proposé par le service universitaire d'activité sportive pour cette année.
- **Adresse de départ :** http://www.univ-tln.fr/
- **Adresse de réponse**
  - (1) http://sport.univ-tln.fr/aviron.php
  - (2) http://sport.univ-tln.fr/documents/administratif/plaquette/Plaquettew-2012-2013.pdf sur la page: http://sport.univ-tln.fr/index.php

Tâche 9 : Recherche de bibliothèque (3) sur le site du Sud Toulon-Var (3)

- **Consigne :** Trouver les horaires d'ouverture/fermeture de la bibliothèque de Droit de Draguignan.

- **Adresse de départ :** http://www.univ-tln.fr/
- **Adresse de réponse**
  - (1) http://bu.univ-tln.fr/medias/medias.aspx?INSTANCE=exploitation&PORTAL_ID=portal_model_instance__horaires.xml&SYNCMENU=UTLV_HORAIRES&VIEW=HOME

Tâche 10 : Recherche de formation (1) sur le site de Brest (4)

- **Consigne :** Trouver le programme d'enseignements avec le volume horaire des cours pour le Master en Sciences Biologiques Marines, spécialité Ecosystèmes marins.
- **Adresse de départ :** http://www.univ-brest.fr/
- **Adresse de réponse**
  - (1) http://formations.univ-brest.fr/fiche/FR_RNE_0290346U_PROG20215/FR_RNE_0290346U_PROG20237/programme

Tâche 11 : Recherche de sport (2) sur le site de Brest (4)

- **Consigne :** Trouver le(s) jour(s) de séance de musculation à Quimper proposé par le service universitaire d'activité sportive pour cette année.
- **Adresse de départ :** http://www.univ-brest.fr/
- **Adresse de réponse**
  - (1) http://www.univ-brest.fr/digitalAssets/7/7448_ACTIVITEQUIMPER.pdf sur la page : http://www.univ-brest.fr/menu/campus/Sport/Le-SUAPS/
  - (2) http://www-tmp.univ-brest.fr/digitalAssetsUBO/9/9254_1-ACTIVITE_S_SPORTIVES_QUIMPER_1er_sem_2012web-1.pdf sur la page: http://www.univ-brest.fr/SUAPS/Les+activit%C3%A9s+du+SUAPS

Tâche 12 : Recherche de bibliothèque (3) sur le site de Brest (4)

- **Consigne :** Trouver les horaires d'ouverture/fermeture de la bibliothèque de la Pérouse.
- **Adresse de départ :** http://www.univ-brest.fr/
- **Adresse de réponse**
  - (1) http://wwz.ifremer.fr/blp/Services/Informations-pratiques

Tâche 13 : Recherche de formation (1) sur le site de Toulouse Capitole (5)

- **Consigne :** Trouver le programme d'enseignements avec le volume horaire des cours pour le Master 1 en Mathématique, Mention Economie et Statistique.
- **Adresse de départ :** http://www.univ-tlse1.fr/

- **Adresse de réponse**
  - (1) http://www.ut-capitole.fr/formations/nos-diplomes/masters/master-m1-mention-economie-et-statistique--310264.kjsp?RH=ZYZYZYZYZYZYZYZYZYZYZY&ONGLET=3

Tâche 14 : Recherche de sport (2) sur le site de Toulouse Capitole (5)

- **Consigne :** Trouver le(s) jour(s) de séance d'escalade proposé par le service universitaire d'activité sportive pour cette année.
- **Adresse de départ :** http://www.univ-tlse1.fr/
- **Adresse de réponse**
  - (1) http://www.ut-capitole.fr/vie-du-campus/activites-physiques-et-sportives/pratiquer-une-activite/daps-escalade-306624.kjsp?RH=1319206844507

Tâche 15 : Recherche de bibliothèque (3) sur le site de Toulouse Capitole (5)

- **Consigne :** Trouver les horaires d'ouverture/fermeture de la bibliothèque Garrigou.
- **Adresse de départ :** http://www.univ-tlse1.fr/
- **Adresse de réponse**
  - (1) http://www.ut-capitole.fr/recherche/bibliotheques/informations-pratiques/bu-acces-horaires-304367.kjsp?RH=ZYZYZYZYZYZYZYZYZYZYZY

ANNEXE 3 – MATERIEL DE L'ETUDE 3

A. Consigne et grilles de notation pour l'évaluation experte des enregistrements faciaux et d'échanges textuels

Evaluation experte d'enregistrements faciaux et d'échanges textuels
Consigne de l'étude

Cette étude vise à estimer la validité et la fiabilité de deux types de donnée : les enregistrements faciaux de webcam et les échanges textuels extrait de chats. De manière plus spécifique, cette étude vise à estimer leurs capacités à renseigner de l'immersion et du plaisir des auditeurs lors d'une activité de téléconférence.

Pour cela, plusieurs vidéos et textes numérotés vous seront présentés. Pour chacun(e)s, il vous sera demandé d'évaluer, selon vos propres capacités et votre intuition, quel est pour vous le niveau d'immersion et de plaisir de l'auditeur en question (1 est le minimum et 6 le maximum). De plus, il vous sera demandé, pour chacune de vos estimation, de leur donner un indice de confiance (1 = Incertain, 6 = Certain).

La manière dont vous parcourez les données sont laissé à votre discrétion (regarder entièrement les vidéos, seulement des extraits, les premières minutes, …). De plus, vos techniques, méthodes ou impressions sur l'exercice ne doivent pas être communiqué aux autres évaluateurs, sous peine de les influencer dans leurs jugements. De mêmes, il convient de ne pas diffuser, montrer les vidéos à un tiers ou à un des auditeurs en questions, afin de respecter la clause d'anonymat et de bonne utilisation de données. Enfin, il est interdit de se faire aider par un autre évaluateur, auditeurs ou autres tiers : cette étude ne fournira de résultats valables que si l'exercice est réalisé de manière individuelle.

Merci pour le temps et le sérieux que vous accorderez à la réalisation de cette étude, estimée à 1-2 jours de travail plein.

1. Tableau d'observation vidéos (Webcam)

	Plaisir						Indice de Confiance						Immersion						Indice de Confiance					
Video 1	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 2	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 3	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 4	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 5	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 7	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 8	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 9	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 10	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 11	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 12	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 13	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 14	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 15	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 16	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 17	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 18	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 19	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 20	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 21	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 22	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 23	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 24	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 25	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 26	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 27	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 28	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 29	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Video 30	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6

2. Tableau d'évaluation des échanges textuels (l'interlocuteur à évaluer est <u>en gras</u> dans le texte) 1/2

	Plaisir						Indice de Confiance						Immersion						Indice de Confiance					
Texte 1	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 2	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 3	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 4	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 5	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 7	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 8	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 9	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 10	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 11	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 12	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 13	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 14	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 15	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 16	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 17	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 18	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 19	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 20	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 21	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 22	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 23	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 24	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 25	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 26	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 27	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 28	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 29	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 30	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6

3. Tableau d'évaluation des échanges textuels (l'interlocuteur à évaluer est <u>en gras</u> dans le texte) 2/2

	Plaisir						Indice de Confiance						Immersion						Indice de Confiance					
Texte 31	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 32	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 33	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 34	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 35	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 36	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 37	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 38	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 39	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 40	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 41	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 42	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 43	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 44	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 45	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 46	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 47	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 48	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 49	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 50	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 51	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 52	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 53	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 54	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 55	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 56	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 57	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 58	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 59	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Texte 60	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6

B. Consigne utilisateur pour l’expérience 1

Consignes générales

« Bonjour, merci d'avoir accepté de participer à cette étude.

Elle est composée de trois étapes :

1ère étape : vous serez invités à vous détendre et vous relaxer durant 5 minutes

2ème étape : vous devrez visionner très attentivement une première vidéo via l'interface d'une nouvelle application. Des questions relatives au contenu de cette vidéoconférence vous seront posées à l'issu du test. D'autres personnes visionneront cette vidéoconférence en même temps que vous. Vous pourrez interagir avec elles par le biais du chat intégré. Lorsque la vidéo sera visionnée, vous devrez répondre à un questionnaire dont la première partie portera sur le contenu de la vidéo, puis dont la seconde partie portera sur l’expérience que vous venez de vivre en tant qu’utilisateur.

3ème étape : vous devrez visionner très attentivement une seconde vidéo via l'interface d'une nouvelle application. Des questions relatives au contenu de cette vidéoconférence vous seront à nouveau posées à l'issu du test. D'autres personnes visionneront également cette vidéoconférence en même temps que vous. Vous pourrez toujours interagir avec elles par le biais du chat intégré. Lorsque la vidéo sera visionnée, vous devrez répondre à un dernier questionnaire dont la première partie portera sur le contenu de la vidéo, puis dont la seconde partie portera sur l’expérience que vous venez de vivre en tant qu’utilisateur. »

Consignes relaxation

« Nous allons commencer par vous demander de vous détendre...

Installez-vous confortablement sur la chaise, détendez vos muscles, respirez à fond, et faites le vide dans votre tête.

Votre main gauche doit être posée à plat, paume vers le haut, sur le support qui est mis à votre disposition à votre gauche.

Nous vous offrons cinq minutes de relaxation, profitez-en ! »

Consignes vidéo

« A présent, je vais vous demander d'assister à une vidéoconférence à distance, via une nouvelle interface.

Trois autres personnes assisteront à cette vidéoconférence en même temps que vous.

Il vous sera possible de consulter leurs profils en survolant leurs portraits avec la souris.

Il vous sera également possible d'échanger quelques propos avec eux via un chat multiple. N'hésitez pas à leur poser des questions ou à leur émettre des commentaires si vous le souhaitez.

***RAPPEL :** Votre main gauche doit être posée à plat, paume vers le haut, sur le support qui est mis à votre disposition à votre gauche. »*

C. Protocole opératoire et consigne utilisateur pour l'expérience 2

1. *Accueil du sujet dans la salle*
2. *Installation du sujet*

« Bonjour, merci d'avoir accepté de participer à cette étude, je vous en prie prenez place… »

[Indiquer l'emplacement du siège du post de test]

« Je vais vous décrire brièvement le déroulement général de ce test, nous en avons pour environ 40 minutes.

Le but de cette expérimentation est de vous faire tester une simulation de vidéoconférence dont le thème porte sur le sommeil. Vous aurez à visionner deux courtes vidéoconférences et à remplir un questionnaire qui s'affichera à la fin de chacune d'elles. Pour vous mettre en situation la plus réelle possible, la simulation intègre la présence de personnes fictives à distance qui interviendront dans l'application par le biais d'un chat (vous ne pourrez donc pas interagir dessus). Par contre vous pourrez être actif et interagir autrement. Des signaux visuels vous guideront, vous êtes libre de les suivre ou non selon votre envie…

Avant de commencer je vais vous poser quelques rapides questions pour établir votre profil »

3. *Remplissage du tableau Excel en indiquant les réponses de chaque sujet*

« Mon objectif dans ce test est de relever des mesures physiologiques en rapport à l'immersion. Pour cela, je vais vous fixer des électrodes sur votre main gauche si vous êtes droitier (et contraire), permettant de relever la conductance cutanée de votre peau. Ensuite, je vais fixer un capteur sur le lobe de votre oreille pour avoir une indication de votre rythme cardiaque pendant l'expérience. Une troisième mesure sera effectuée en parallèle, la capture du mouvement oculaire de vos yeux, qui sera faite par le biais d'un eye-tracker, l'écran situé en face de vous. Toutes ces mesures me permettront d'analyser votre état émotionnel pendant que vous observerai les vidéos. Je pourrais vous montrer les courbes physiologiques à la fin de l'expérience si cela vous intéresse.

Voici les deux électrodes que je vais vous fixer sur la main, veuillez placer votre main paume vers le haut, il est important d'éviter le plus possible de bouger cette main (sans aller jusqu'au crampe bien sûre) pendant l'observation des vidéos.

Voici le capteur pour le rythme cardiaque, je vais le placer sur votre lobe d'oreille gauche. »

4. *Branchement des capteurs GSR sur deux phalanges de la main*
5. *Vérification de la posture correcte de la main (paume vers le haut)*
6. *Branchement du capteur de PPG sur le lobe de l'oreille*

« C'est bon ? Cela ne pince pas trop fort ?

Etes-vous confortablement installé ?

Concernant le suivi oculaire, un calibrage de l'eye-tracker doit être fait, quand vous verrez apparaitre un point rouge sur l'écran, veuillez le suivre jusqu'au bout.

Toutes les étapes seront affichées au fur et à mesure sur l'écran. »

7. *Vérification ou réglage de l'assise sur le fauteuil pour être bien face à l'écran de l'eye-tracker*

« Nous allons commencer les mesures, avant que la première vidéoconférence ne se lance, je vais vous demander de vous relaxer quelques minutes sans parler.

C'est bon ? Êtes-vous prêt ? Vous pouvez alors mettre le casque audio, nous allons démarrer.

C'est parti ! »

8. *Mise en place du casque audio pour l'écoute des vidéos*
9. *Etablir le calibrage de l'eye tracking*
10. *Etablir la baseline de la GSR (calibrage) : quelques minutes de repos*
11. *Lancement de la première vidéoconférence (cf. contrebalancement)*
12. *Présentation au sujet du premier questionnaire d'immersion à remplir*
13. *Diffusion de la deuxième vidéoconférence*
14. *Présentation au sujet du deuxième questionnaire d'immersion à remplir*
15. *Etablir l'amplitude maximale du GSR un mini jeu vidéo*

« Voilà, l'expérience est terminée, merci d'avoir participer, je peux vous montrer les graphs correspondant à votre niveau de conductance cutanée et rythme cardiaque si cela vous intéresse ? »

[Si le sujet est intéressé, consacrer cinq minute à la présentation et à l'explication des données recueillies.]

« Avez-vous des commentaires à émettre par rapport aux vidéos ? Qu'avez-vous pensé de l'expérience en générale ? »

[Développer les thèmes de l'attention et de la satisfaction perçue, de l'intérêt porté et du niveau de difficulté de compréhension ressenti. Noter l'ensemble des remarques pertinentes dans la colonne commentaire de la grille d'utilisation]

« Au revoir ! »

11. *Remerciement, recueil de commentaire et échange avec le sujet, présentation de la courbe*
12. *GSR et PPG en fonction de l'intérêt du sujet.*
13. *Fin de l'expérimentation*

D. Debrief des experts sur la méthode et les indices utilisés pour évaluer les vidéos et les traces écrites

Expert 1

Comment avez-vous analysé les vidéos ?

Je me suis aidé d'une grille permettant de compter la fréquence de certain comportements de l'utilisateur. J'ai ainsi pu compter les soupirs, clignements des yeux, fixations hors champs, bayements, sourires, agitations et regards dans le chat pour chaque vidéo. J'ai regardé les vidéos en entier sauf pour certaine personnes très calmes pour lesquels je m'autorisais des petits sauts en avant de 5 seconds environs. Mon temps d'analyse a été de 5H environ.

Quels ont été pour vous les indicateurs d'immersion dans les vidéos ?

Les indicateurs de faible immersion sont : un nombre élevé de bayements, d'épisodes d'agitations, de clignement des yeux (fatigue), et la durée de regards hors champs ou sur le chat. Les indicateurs d'immersion élevée sont le nombre de sourires long ainsi que de longues fixations sur la vidéo.

Quels ont été pour vous les indicateurs de plaisir dans les vidéos ?

Le nombre de sourires (surtout long) pour le plaisir. Les expressions faciales, comportement et les positions du corps traduisant l'ennui pour le déplaisir (tronche pas possible, soupirs, regard hors champs ou dans le vide).

Quels étaient les facteurs qui jouaient sur la confiance de vos estimations ?

Il s'agissait du nombre d'indice. De plus, il est plus facile de repérer un état de non immersion que d'immersion (moins d'indices comportementaux).

Comment avez-vous analysé les échanges textuels ?

J'ai lu chacun des échanges en entier. Mon temps d'analyse a été de 2 heures environ.

Quels ont été pour vous les indicateurs d'immersion dans les échanges textuels ?

Cela a été surtout ce qui tournait autour de l'acte d'écriture. C'est le nombre de commentaires et de questions, qui démontre que la personne suit et s'intéresse. Ils ne doivent pas trop être élevés non plus, ce qui démontrerait que la personne ne prête plus attention à la présentation. Les commentaires, positifs ou négatifs, peuvent également être des indicateurs de l'immersion.

Quels ont été pour vous les indicateurs de plaisir dans les échanges textuels ?

Cela a été surtout sur le contenu de ce que la personne écrivait. Le plaisir est le déplaisir transparait principalement dans la valence des commentaires. Toutefois, la satisfaction des utilisateurs peut également transparaitre dans le niveau d'implication du sujet (commentaire, questions, …)

Quels étaient les facteurs qui jouaient sur la confiance de vos estimations ?

Elle varie selon la quantité d'échanges disponible. C'est un exercice plus difficile que dans le cas des vidéos malgré le fait que cela soit plus facile pour l'estimation du plaisir.

Expert 2

Comment avez-vous analysé les vidéos ?

J'ai regardé les 30 premières secondes de chaque vidéo puis avancé de 3 secondes à la fois, image par image, pour bien voir les changements de mouvement et d'expression faciale. J'ai mis environ 5 heures pour tout analyser.

Quels ont été pour vous les indicateurs d'immersion dans les vidéos ?

Les indicateurs de bonne immersion ont été pour moi une position du corps orienté vers l'écran, pas/de mouvements du corps et/ou des yeux, une expression faciale concentré. Les indicateurs de faibles immersions ont été pour moi les expressions faciales de déplaisir/tic faciaux, les regards hors écrans, le corps ou la tête tournée ailleurs. Ces indicateurs ne sont pas à évaluer au tout début de la vidéo car la personne ne s'est pas encore familiarisée avec le media.

Quels ont été pour vous les indicateurs de plaisir dans les vidéos ?

Les sourires, la joie transparaissant sur le visage, la personne qui se rapproche du support. Le déplaisir est indiqué par un regard en dehors de la vidéo.

Quels étaient les facteurs qui jouaient sur la confiance de vos estimations ?

Elle dépend de l'expressivité des gens, voire de sa personnalité (car il y a des personnes qui sourient tout le temps). Pour cela les premières secondes de vidéo et les premières réactions sont très instructives.

Comment avez-vous analysé les échanges textuels ?

J'ai lu les échanges des auditeurs en premier, puis des autres auditeurs quand j'avais une incompréhension. J'ai mis environ 5 heures pour tout analyser.

Quels ont été pour vous les indicateurs d'immersion dans les échanges textuels ?

Les indicateurs de bonne immersion ont été pour moi quand les auditeurs n'échangeaient pas/peu de texte ou demandaient aux autres d'arrêter de parler (sauf au début, car ce n'est pas représentatif.). Les indicateurs de faible immersion ont été pour moi que la personne parlait beaucoup et sur des sujets n'ayant rien à voir.

Quels ont été pour vous les indicateurs de plaisir dans les échanges textuels ?

Les indicateurs de plaisir ont été pour moi les même que l'immersion (car contribuant au plaisir/déplaisir), avec en plus l'explicitation de son intérêt/désintérêt. Il y a aussi les moments où l'utilisateur formule des commentaires en rapport avec le sujet (ah je ne savais pas…), ce qui prouve son investissement. En indicateur de plaisir négatif, il y a les discussions hors sujets

Quels étaient les facteurs qui jouaient sur la confiance de vos estimations ?

J'étais moins confiant quand il y avait peu d'échanges et plus confiant quand la personne explicitait directement son intérêt pour la vidéo.

Expert 3

Comment avez-vous analysé les vidéos ?

J'ai parcouru rapidement toutes les vidéos, puis j'ai repéré les pairs pour voir les différences. J'ai regardé la moitié des vidéos en vitesse normale, puis le reste en vitesse accélérée pour confirmer mon évaluation. J'ai mis environ 4 heures pour tout analyser.

Quels ont été pour vous les indicateurs d'immersion dans les vidéos ?

Les indicateurs de bonne immersion ont été pour moi principalement le regard. C'est le nombre d'aller retours entre le chat et la vidéo (regard à gauche : chat, et à droite : vidéo) et également la proportion. En indicateur négative, il y avait également les regards hors champs.

Quels ont été pour vous les indicateurs de plaisir dans les vidéos ?

Les indicateurs de plaisir ont été pour moi les expressions faciales et corporelles. Il y avait surtout des regards neutres ou négatifs, les rires étant surtout dus aux discussions sur le chat. En positif donc, c'était un regard neutre et droit. Pour le négatif, c'était une position de travers ou la tête penchée.

Quels étaient les facteurs qui jouaient sur la confiance de vos estimations ?

C'était plus simple quand la personne regardait hors champs ou au contraire superfixait l'écran, je savais que c'était négatif. Par contre, dans le cas d'une immersion forte, il y a moins de signe, donc j'étais moins sûr.

Comment avez-vous analysé les échanges textuels ?

Au début je lisais tous les échanges. Puis, après avoir repéré les phrases modèles du Wizard of Oz, je ne lisais plus que les répliques des utilisateurs avec le contexte. J'ai mis environ 3 heures pour tout analyser.

Quels ont été pour vous les indicateurs d'immersion dans les échanges textuels ?

Les indicateurs de bonne immersion ont été pour moi quand les auditeurs n'échangeaient pas ou peu, ou alors seulement sur le contenu de la présentation. Les indicateurs de faible immersion ont été pour moi que la personne parlait trop et était hors sujets.

Quels ont été pour vous les indicateurs de plaisir dans les échanges textuels ?

Les indicateurs de plaisir ou de déplaisir ont été pour moi les verbalisations de l'utilisateur qui l'exprimait clairement. Si l'utilisateur ne s'exprimait pas à ce sujet, je notais 3 ou 4, avec une confiance nulle

Quels étaient les facteurs qui jouaient sur la confiance de vos estimations ?

J'étais moins confiant quand je notais des personnes immergées (car moins d'indices)

Expert 4

Comment avez-vous analysé les vidéos ?

J'ai regardé toutes les vidéos en entier. Mon temps d'analyse a été de 4H environs.

Quels ont été pour vous les indicateurs d'immersion dans les vidéos ?

Les indicateurs de faible immersion sont nombreux. Il s'agit des mouvements de corps intensifs (signe d'une faible concentration) et du temps passé sur le chat (ou du nombre d'aller-retour du regard). Les indicateurs d'immersion élevés sont un regard fixe, une posture concernée (en avant, se tenant le menton, signe faciaux d'acquiescement ou de réflexion).

Quels ont été pour vous les indicateurs de plaisir dans les vidéos ?

Pour le positif, les expressions faciales, à travers les sourires, les yeux. Pour le négatif, les bâillements, l'air de ne pas être intéressé (yeux réprobateurs, regards passifs).

Quels étaient les facteurs qui jouaient sur la confiance de vos estimations ?

Il s'agissait du nombre de signaux. De plus, il est plus facile de détecter un état de déconcentration que l'inverse.

Comment avez-vous analysé les échanges textuels ?

J'ai lu les échanges en gras (utilisateurs) et ce qu'il y avait autour. Je regardais aussi la taille globale des échanges. Mon temps d'analyse a été de 4H environ.

Quels ont été pour vous les indicateurs d'immersion dans les échanges textuels ?

Pour les indices positifs, c'était les remarques constructives en rapport avec la vidéo (compréhension, réponses, positionnement), des déclarations positives sur son état d'intérêt (ex : ça m'intéresse) ou alors au contraire, peu d'échanges. Les indices de faible immersion dans les textes étaient le nombre d'hors sujet, des déclarations négatives sur son état d'intérêt (ça me soûle)

Quels ont été pour vous les indicateurs de plaisir dans les échanges textuels ?

Les indicateurs de plaisir était les smileys et les expressions ayant une valence positive (ex : cool, bien). Celles de déplaisir étaient les expressions négatives (ex : je m'ennuis, …)

Quels étaient les facteurs qui jouaient sur la confiance de vos estimations ?

Le nombre et la fréquence des indices impactent la confiance. C'est beaucoup plus difficile d'estimer quand il y a peu d'indices.

Expert 5

Comment avez-vous analysé les vidéos ?

J'ai regardé toutes vidéos en entier. J'ai mis environ 5 heures pour tout analyser.

Quels ont été pour vous les indicateurs d'immersion dans les vidéos ?

Les indicateurs de bonne immersion ont été pour moi une position droite et vers l'écran, avec expression faciale concentré. Les indicateurs de faible immersion les regards hors écrans ou distraits (beaucoup de mouvements).

Quels ont été pour vous les indicateurs de plaisir dans les vidéos ?

Les sourires principalement et pour le déplaisir, un regard passif ou fuyant.

Quels étaient les facteurs qui jouaient sur la confiance de vos estimations ?

Le nombre d'indice, positif ou négatifs, mais cela dépend aussi de l'expressivité des gens.

Comment avez-vous analysé les échanges textuels ?

J'ai lu tous les échanges des auditeurs en premier. J'ai mis environ 4 heures pour tout analyser.

Quel ont été pour vous les indicateurs d'immersion dans les échanges textuels ?

Les indicateurs de bonne immersion ont été pour moi quand les utilisateurs n'échangeaient pas/peu ou déclaraient explicitement être intéressé par le sujet. Les indicateurs de faible immersion ont été pour moi la fréquence d'hors sujets.

Quels ont été pour vous les indicateurs de plaisir dans les échanges textuels ?

Les indicateurs de plaisir ont été principalement les expressions d'intérêt/désintérêt.

Quels étaient les facteurs qui jouaient sur la confiance de vos estimations ?

J'étais moins confiant quand il y avait peu d'échanges et plus confiant quand la personne explicitait directement son intérêt/désintérêt

BIBLIOGRAPHIE

Abowd, G. D., Atkeson, C. G., Bobick, A. F., Essa, I. A., MacIntyre, B., Mynatt, E. D., & Starner, T. E. (2000). Living laboratories: the future computing environments group at the Georgia Institute of Technology. In *Extended Abstracts on Human Factors in Computing Systems (CHI EA '00)* (pp. 215–216). New York: ACM Press. doi:10.1145/633292.633416

Ackerman, M. S., Erickson, T., Halverson, C. A., & Kellogg, W. A. (2008). *Resources, Co-Evolution and Artifacts: Theory in CSCW*. London, UK: Springer-Verlag London Limited.

Adam, A. R., Mallan, K. M., & Lipp, O. V. (2009). The effect of emotional and attentional load on attentional startle modulation. *International Journal of Psychophysiology*, *74*(3), 266–273. doi:10.1016/j.ijpsycho.2009.09.011

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, *17*(6), 734–749. doi:10.1109/TKDE.2005.99

Adriaans, P., & van Benthem, J. (2008a). Introduction: Information is what information does. In P. Adriaans & J. van Benthem (Eds.), *Philosophy of information*. UK: North-Holland.

Adriaans, P., & van Benthem, J. (2008b). *Philosophy of Information*. UK: Elsevier B.V.

Aftanas, L., & Golocheikine, S. A. (2001). Human anterior and frontal midline theta and lower alpha reflect emotionally positive state and internalized attention: high-resolution EEG investigation of meditation. *Neuroscience Letters*, *310*(1), 57–60.

Agarwal, R., & Karahanna, E. (2000). Time Flies When You're Having Fun: Cognitive Absorption and Beliefs about Information Technology Usage. *MIS Quarterly*, *24*(4), 665. doi:10.2307/3250951

Agliati, A., Mantovani, F., Realdon, O., Confalonieri, L., & Vescovo, A. (2006). Multimodal Temporal Patterns for the Analysis of User 's Involvement in Affective Interaction with Virtual Agents. In G. Riva, M. T. Anguera, B. K. Wiederhold, & F. Mantovani (Eds.), *From Communication to Presence: Cognition, Emotions and Culture towards the Ultimate Communicative Experience*. Amsterdam: IOS Press.

Aguinis, H., & Lawal, S. O. (2012). Conducting field experiments using eLancing's natural environment. *Journal of Business Venturing*, *27*(4), 493–505. doi:10.1016/j.jbusvent.2012.01.002

Ahlstrom, U., & Friedman-Berg, F. J. (2006). Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics*, *36*(7), 623–636. doi:10.1016/j.ergon.2006.04.002

Ajanki, A., Hardoon, D. R., Kaski, S., Puolamäki, K., & Shawe-Taylor, J. (2009). Can eyes reveal interest? Implicit queries from gaze patterns. *User Modeling and User-Adapted Interaction*, *19*(4), 307–339. doi:10.1007/s11257-009-9066-4

Aladwani, A. M., & Palvia, P. C. (2002). Developing and validating an instrument for measuring user-perceived web quality. *Information and Management*, *39*(6), 467–476. doi:10.1016/S0378-7206(01)00113-6

Alavi, M., & Leidner, D. E. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly*, *25*(1), 107–136. Retrieved from http://www.jstor.org/stable/3250961

Alben, L. (1996). Quality of experience: defining the criteria for effective interaction design. *Interactions*, *3*(3), 11–15. doi:10.1145/235008.235010

Alexa Internet Inc. (2013). Alexa top 500 global sites. Retrieved from http://www.alexa.com/topsites

Aley, E., Cooper, T., Graeber, R., Kerne, A., Overby, K., & Toups, Z. O. (2005). Censor chair: Exploring Censorship and Social Presence through Psychophysiological Sensing. In *Proceedings of the 13th annual ACM international conference on Multimedia - MULTIMEDIA '05* (p. 922). New York, New York, USA: ACM Press. doi:10.1145/1101149.1101345

Alkhattabi, M., Neagu, D., & Cullen, A. (2011). Assessing information quality of e-learning systems: a web mining approach. *Computers in Human Behavior*, *27*(2), 862–873. doi:10.1016/j.chb.2010.11.011

Alonso-Ríos, D., Vázquez-García, A., Mosqueira-Rey, E., & Moret-Bonillo, V. (2009). Usability: A Critical Analysis and a Taxonomy. *International Journal of Human-Computer Interaction*, *26*(1), 53–74. doi:10.1080/10447310903025552

Altshuller, G. S. (1999). *The Innovation Algorithm: TRIZ, Systematic Innovation and Technical Creativity*. Worcester, MA: Technical Innovation Center.

Amabile, T. M. (1996). *Creativity in context*. Boulder, Colorado: Westview Press.

Amabile, T. M. (1996). *Creativity and innovation in organizations*. Boston, MA.

Amabile, T. M. (1998). How to kill creativity ? *Harvard Business Review*, *76*(5), 76–87.

Ambler, S. (2002). *Agile modeling: effective practices for extreme programming and the unified process*. London: Wiley.

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME), &. (2002). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Anastassova, M., Mégard, C., & Burkhardt, J. (2007). Prototype evaluation and user-needs analysis in the early design of emerging technologies. In J. A. Jacko (Ed.), *Human-Computer Interaction. Interaction Design and Usability* (Vol. 4550, pp. 383–392). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-73105-4

Anderson, C. (2009). *Free: How today's smartest businesses profit by giving something for nothing*. New York: Hyperion.

Andresen, M. A. (2009). Asynchronous discussion forums : success factors , outcomes , assessments , and limitations. *Educational Technology & Society*, *12*(1), 249–257.

Annells, M. (2011). Grounded Theory Method: Philosophical Perspectives, Paradigm of Inquiry, and Postmodernism. In S. Delamont & P. Atkinson (Eds.), *SAGE qualitative research methods* (pp. 380–391). London: SAGE. doi:10.4135/9780857028211

ANSI. (2001). Common industry format for usability test reports (ANSI-NCITS 354-2001). Washington, DC: American National Standards Institute.

Anttonen, J., Surakka, V., & Koivuluoma, M. (2009). Ballistocardiographic Responses to Dynamic Facial Displays of Emotion While Sitting on the EMFi Chair. *Journal of Media Psychology: Theories, Methods, and Applications*, *21*(2), 69–84. doi:10.1027/1864-1105.21.2.69

Antunes, P., & Costa, C. J. (2003). Perceived Value: A Low-Cost Approach to Evaluate Meetingware. In *Proceedings of CRIWG'03, Autrans, France, Lecture Notes in Computer Science 2806* (pp. 109–125). Springer Berlin Heidelberg. doi:10.1.1.65.700

Antunes, P., Herskovic, V., Ochoa, S. F., & Pino, J. a. (2012). Structuring dimensions for collaborative systems evaluation. *ACM Computing Surveys*, *44*(2), 1–28. doi:10.1145/2089125.2089128

Antunes, P., Ramires, J., & Respício, A. (2006). Addressing the Conflicting Dimension of Groupware: A Case Study in Software Requirements Validation. *Computing and Informatics*, *25*(6), 523–546.

Arthur, B. (1989). Competing technologies, increasing returns, and lock-in by historical events. *The Economic Journal*, *99*, 116–131.

Arthur, C. (2009, June). The breakneck race to build an application to crowdsource MPs' expenses. *The Guardian*. Retrieved from http://www.guardian.co.uk/

Ashenfelter, O., & Quandt, R. (1999). Analyzing wine tasting statistically. *Chance*, *12*(3), 16–20.

Asteriadis, S., Karpouzis, K., & Kollias, S. (2009). Feature Extraction and Selection for Inferring User Engagement in an HCI Environment. In J. A. Jacko (Ed.), *Human-Computer Interaction. New Trends* (Vol. 5610, pp. 22–29). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-02574-7

Atallah, L., & Yang, G.-Z. (2009). The use of pervasive sensing for behaviour profiling — a survey. *Pervasive and Mobile Computing*, *5*(5), 447–464. doi:10.1016/j.pmcj.2009.06.009

Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, *16*(6), 345–379. doi:10.1007/s00530-010-0182-0

Bacci, S., Bartolucci, F., & Gnaldi, M. (2014). Multidimensional latent class item response models for binary and ordinal polytomous items: theory and application with R software. *Communications in Statistics - Theory and Methods*, *43*(4), 787–800. doi:10.1080/03610926.2013.827718

Bachelard, G. (1970). *La formation de l'esprit scientifique*. Paris: Vrin.

Backs, R. W., & Seljos, K. A. (1994). Metabolic and cardiorespiratory measures of mental effort : The effects of level of difficulty in a working memory task. *International Journal of Psychophysiology*, *16*(1), 57–68.

Baeza-Yates, R., & Pino, J. A. (1997). A first step to formally evaluate collaborative work. In *Proceedings of the international ACM SIGGROUP conference on Supporting group work : the integration challenge the integration challenge - GROUP '97* (pp. 56–60). New York, New York, USA: ACM Press. doi:10.1145/266838.266860

Baeza-Yates, R., & Pino, J. A. (2006). Towards formal evaluation of collaborative work. *Information Research*, *11*(4). Retrieved from http://informationr.net/ir/11-4/paper271.html

Bailenson, J. N., Pontikakis, E. D., Mauss, I. B., Gross, J. J., Jabon, M. E., Hutcherson, C. a. C., … John, O. (2008). Real-time classification of evoked emotions using facial feature tracking and physiological responses. *International Journal of Human-Computer Studies*, *66*(5), 303–317. doi:10.1016/j.ijhcs.2007.10.011

Baker, K., Greenberg, S., & Gutwin, C. (2001). Heuristic Evaluation of Groupware Based on the Mechanics of Collaboration. In M. R. Little & L. Nigay (Eds.), *Proceedings of the 8th IFIP international Conference on Engineering For Human- Computer interaction, Lecture Notes In Computer Science 2254* (Vol. 2254, pp. 123–139). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/3-540-45348-2_14

Baker, K., Greenberg, S., & Gutwin, C. (2002). Empirical development of a heuristic evaluation methodology for shared workspace groupware. In *Proceedings of the 2002 ACM conference on Computer supported*

cooperative work - CSCW '02 (p. 96). New York, New York, USA: ACM Press. doi:10.1145/587078.587093
Bakker, S., Markopoulos, P., & de Kort, Y. (2008). OPOS: an observation scheme for evaluating head-up play. In *Proceedings of the 5th Nordic Conference on Human-Computer interaction: Building Bridges (NordiCHI '08)* (pp. 18–22). New York: ACM Press. doi:10.1145/1463160.1463165
Baldo, B. A., & Kelley, A. E. (2007). Discrete neurochemical coding of distinguishable motivational processes : Insights from nucleus accumbens control of feeding. *Psychopharmacology*, *191*, 439–459.
Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An Empirical Evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, *24*(6), 574–594. doi:10.1080/10447310802205776
Bannister, D., & Fransella, F. (1985). *Inquiring man. 3rd Edition*. London, UK: Routledge.
Barcenilla, J., & Bastien, J.-M.-C. (2009). L'acceptabilité des nouvelles technologies : quelles relations avec l'ergonomie, l'utilisabilité et l'expérience utilisateur ? *Le Travail Humain*, *72*(4), 311. doi:10.3917/th.724.0311
Bardini, T. (2000). *Bootstrapping: Douglas Engelbart, coevolution, and the origins of personal computing*. Stanford, Calif: Stanford University Press.
Bargas-Avila, J. A., & Hornbæk, K. (2011). Old wine in new bottles or novel challenges ? A critical analysis of empirical studies of user experience. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11* (p. 2689). New York, New York, USA: ACM Press. doi:10.1145/1978942.1979336
Barney, J. B. (1991). Firm Resources and Sustained Competitive Advantage. *Journal of Management*, *17*(1), 99–120.
Barrère, G., Sloim, E., Bastien, C., & Mazzone, E. (2012). *Card sorting: Ne perdez plus vos utilisateurs !*. Paris: Eyrolles.
Bartholomew, D. J. (2006). *Measurement: Sage Benchmarks in Social Research Methods, vols. I–IV*. London: Sage.
Bartlett, F. C. (1962). The future of ergonomics. *Ergonomics*, *5*, 505–511.
Bassani, M., Sbalchiero, S., Ben Youssef, K., & Magne, S. (2010). *Brand Design : Construire la personnalité d'une marque gagnante*. Bruxelles: De Boeck.
Bastien, J.-M.-C., Brangier, E., Dinet, J., Barcenilla, J., Michel, G., & Vivian, R. (2009). The Expert Community Staff: An innovative method for capturing end-users' needs. In L. Norros, H. Koskinen, L. Salo, & P. Savioja (Eds.), *Designing beyond the product: understanding activity an user experience in ubiquitous environments. European Conference on Cognitive Ergonomics (ECCE'2009)* (pp. 374–379).
Battarbee, K. (2003). Defining Co-Experience. In *Proceedings of DPPI'03* (pp. 109–113). New York, NY, USA: ACM Press.
Bawden, D. (2001). Information overload. *Library & Information Briefings*, *92*, 1–15.
Bawden, D., & Robinson, L. (2008). The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of Information Science*, *35*(2), 180–191. doi:10.1177/0165551508095781
Beaudouin-Lafon, M., & Mackay, W. E. (2009). Prototyping tools and techniques. In A. Sears & J. A. Jacko (Eds.), *Human Computer Interaction : Development Process* (pp. 121–143). New York: CRC Press.
Beck, K. (1999). *Extreme Programming Explained: Embrace Change*. Reading, MA: Addison-Wesley Publishing.
Beguin, P. (2007). Prendre en compte l'activité de travail pour concevoir. *@ctivités*, *4*(2), 107–114.
Bellman, R. E. (1957). *Dynamic programming*. Princeton: Princeton University Press.
Bellucci, A., Malizia, A., Diaz, P., & Aedo, I. (2010). The Anatomy of Web 2 . 0 : The Web as a Platform to Promote Users ' Participation and Collaboration. In F. V. Cipolla-Ficarra (Ed.), *Quality and Communicability for Interactive Hypermedia Systems: Concepts and Practices for Design* (pp. 36–63). IGI Global. doi:10.4018/978-1-61520-763-3.ch003
Benghozi, P.-J., Bureau, S., Massit-Folléa, F., Waroquiers, C., & Davidson, S. (2009). *L'internet des objets*. Paris: Maison des sciences de l'homme.
Benkler, Y. (2006). *The wealth of networks how social production transforms markets and freedom*. New Haven: Yale University Press.
Bennett, C. M., Baird, A. A., Miller, M. B., & Wolford, G. L. (2010). Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon : An Argument for Multiple Comparisons Correction. *Journal of Serendipitous and Unexpected Results*, *1*(1).
Bennett, J. L. (1972). The user interface in interactive systems. *Annual Review of Information Science*, *7*, 159–196.
Bennis, W., & Biederman, P. W. (1998). None of Us Is As Smart As All of Us. *IEEE Computer*, *31*(3), 116–117.
Berg, J. E., Nelson, F. D., & Rietz, T. A. (2008). Prediction market accuracy in the long run. *International Journal of Forecasting*, *24*(2), 283–298. doi:10.1016/j.ijforecast.2008.03.007
Berlyne, D. E. (1960). *Conflict, arousal, and curiosity*. New York, NY, USA: McGraw-Hill.

Berlyne, D. E. (1970). Novelty, complexity and hedonic value. *Perception and Psychophysics*, *8*(279-286).
Bernstein, M. S., Brandt, J., Miller, R. C., & Karger, D. R. (2011). Crowds in two seconds : Enabling realtime crowd-powered interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11* (p. 33). New York, New York, USA: ACM Press. doi:10.1145/2047196.2047201
Berntson, G. G., Cacioppo, J. T., & Quigley, K. S. (1993). Respiratory sinus arrhythmia: Autonomic origins, physiological mechanisms, and psychophysiological implications. *Psychophysiology*, *30*(2), 183–196.
Bevan, N. (1999). Quality in use: Meeting user needs for quality. *Journal of Systems and Software*, *49*(1), 89–96. doi:10.1016/S0164-1212(99)00070-9
Bevan, N. (2008). UX, usability and ISO standards. In *CHI 2008 Workshop on User Experience Evaluation Methods in Product Development*. Florence, Italy.
Beyer, H., & Holtzblatt, K. (1998). *Contextual design: defining customer-centered systems*. San Francisco: Morgan Kaufmann.
Bianchi-berthouze, N. (2013). Understanding the Role of Body Movement in Player Engagement. *Human–Computer Interaction*, *28*(1), 40–75. doi:10.1080/07370024.2012.688468
Bias, R. (1991). Interface-Walkthroughs: efficient collaborative testing. *IEEE Software*, *8*(5), 94–95. doi:10.1109/52.84220
Bias, R. (1994). The pluralistic usability walkthrough: coordinated empathies. In J. Nielsen & R. Mack (Eds.), *Usability inspection Methods* (pp. 63–76). New York: John Wiley & Sons.
Bickerton, D. (1990). *Language and Species*. University of Chicago Press.
Bisset, F. (2012). Making out the process of design. *fergusbisset.com*. Retrieved August 16, 2014, from http://www.fergusbisset.com/2012/09/26/marking-out-the-process-of-design/
Björn-Andersen, N., & Hedberg, B. (1977). Design of information systems in an organizational perspective. In P. C. Nystrom & W. H. Starbuck (Eds.), *Prescriptive models of organizations. TIMS Studies in the Management Sciences, Vol. 5* (pp. 125–142). Amsterdam: North-Holland.
Blanchot, F., & Fort, F. (2007). Coopétition et alliances en R&D. *Revue Française de Gestion*, *33*(7), 163–182. doi:10.3166/rfg.176.163-182
Blank, G., & Reisdorf, B. C. (2012). The participatory web : A user perspective on Web 2.0. *Information, Communication & Society*, *15*(4), 537–554. doi:10.1080/1369118X.2012.665935
Blythe, M. A., Monk, A. F., Overbeeke, K., & Wrigh, P. C. (2003). *Funology : From Usability to Enjoyment*. Dordrecht: Springer Science + Business Media, Inc.
Bocker, M., & Muhlbach, L. (1993). Communicative presence in videocommunications. In *Proceedings of the Human Factors and Ergonomics Society 37th annual meeting* (pp. 249–253). Santa Monica, CA: Human Factors and Ergonomics Society.
Boden, M. A. (2008). Information, computation, and cognitive science. In P. Adriaans & J. van Benthem (Eds.), *Philosophy of Information*. UK: Elsevier B.V.
Boehner, K., Depaula, R., Dourish, P., & Sengers, P. (2007). How emotion is made and measured. *International Journal of Human-Computer Studies*, *65*(4), 275–291. doi:10.1016/j.ijhcs.2006.11.016
Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: capturing life as it is lived. *Annual Review of Psychology*, *54*(1), 579–616. doi:10.1146/annurev.psych.54.101601.145030
Bommer, W. H., Johnson, J. L., Rich, G. A., Podsakoff, P. M., & Mackenzie, S. B. (1995). On the interchangeability of objective and subjective measures of employee performance: a meta-analysis. *Personnel Psychology*, *48*, 587–605.
Boorstin, D. J. (1985). *The discoverers*. New York: Vintage Books.
Borsci, S., Federici, S., Bacci, S., Gnaldi, M., & Bartolucci, F. (2015). Assessing User Satisfaction in the Era of User Experience: Comparison of the SUS, UMUX, and UMUX-LITE as a Function of Product Experience. *International Journal of Human-Computer Interaction*, *31*(8), 484–495. doi:10.1080/10447318.2015.1064648
Borsci, S., Federici, S., & Lauriola, M. (2009). On the dimensionality of the System Usability Scale: a test of alternative measurement models. *Cognitive Processing*, *10*(3), 193–7. doi:10.1007/s10339-009-0268-9
Borsci, S., Macredie, R. D., Barnett, J., Martin, J., Kuljis, J., & Young, T. (2013). Reviewing and Extending the Five-User Assumption: A Grounded Procedure for Interaction Evaluation. *ACM Transactions on Computer-Human Interaction*, *20*(5), 1–23. doi:10.1145/2506210
Bossink, B. A. G. (2002). The development of co-innovation strategies : stages and interaction pattern in interfirm. *R&D Management*, *32*(4), 311–320.
Bouchard, C. (1997). *Modélisation du processus de design automobile. Méthode de veille stylistique au design du composant d'aspect*. Ecole Nationale Supérieure Arts et Métiers. ParisTech.
Bouchard, S., Talbot, J., Ledoux, A., Phillips, J., Cantamasse, M., & Robillard, G. (2009). The Meaning of Being There is Related to a Specific Activation in the Brain Located in the Parahypocampus. In *Proceeding of the Presence 2009 : the 12th Annual International Workshop on Presence*.

Boucsein, W. (1992). *Electrodermal activity*. New York: Plenum Press.
Boulier, D. (2012). L'âge de la prédation. Note de lecture sur L'âge de la multitude de Colin et Verdier aux éditions Armand Colin. *InternetActu*. Retrieved from http://www.internetactu.net/2012/09/07/l%E2%80%99age-de-la-predation/
Boullier, D. (2010a). *Le client du poste téléphonique : archéologie des êtres intermédiaires*. *Débordements. Mélanges pour Michel Callon* (pp. 41–61). Presses de l'Ecole des Mines.
Boullier, D. (2010b). Six recettes pour ne pas innover (et l'inverse). In *Colloque Innovacs*. Grenoble, France.
Boullier, D. (2012). L'âge de la prédation : Note de lecture sur L'âge de la multitude de Colin et Verdier aux éditions Armand Colin. *InternetActu*.
Boullier, D., & Lohard, A. (2012). *Opinion mining et sentiment analysis*. Marseille: OpenEdition Press.
Bourdieu, P. (1979). *La distinction. Critique sociale du jugement*. Paris: Editions de Minuit.
Bower, J. . L., & Christensen, C. M. (1995). Disruptive Technologies : Catching the Wave. *Harvard Business Review*, *73*(1), 43–53.
Boyle, E. a., Connolly, T. M., Hainey, T., & Boyle, J. M. (2012). Engagement in digital entertainment games: A systematic review. *Computers in Human Behavior*, *28*(3), 771–780. doi:10.1016/j.chb.2011.11.020
Brabham, D. C. (2008). Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. *Convergence: The International Journal of Research into New Media Technologies*, *14*(1), 75–90. doi:10.1177/1354856507084420
Bradley, J., & Benyon, D. (2009). Wizard of Oz Experiments for Companions. In *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology (BCS-HCI '09)* (pp. 313–317). Swinton, UK, UK: British Computer Society.
Bradley, J., Reberger, C., Dixit, A., & Gupta, V. (2013). *Internet of Everything: A $4.6 Trillion Public-Sector Opportunity (White Paper)*.
Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59. doi:10.1016/0005-7916(94)90063-9
Bradley, M. M., Miccoli, L., Escrig, M. a, & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*(4), 602–607. doi:10.1111/j.1469-8986.2008.00654.x
Brajnik, G., & Giachin, C. (2014). Using sketches and storyboards to assess impact of age difference in user experience. *International Journal of Human-Computer Studies*, *72*(6), 552–566. doi:10.1016/j.ijhcs.2013.12.005
Brangier, E. (2003). Le concept de « symbiose homme-technologie-organisation ». In N. Delobbe, G. Karnas, & C. Vandenberg (Eds.), *Évaluation et développement des compétences au travail (Volume 3)* (pp. 413–422). UCL: Presses universitaires de Louvain.
Brangier, E., & Bastien, J. M. C. (2006). L'analyse de l'activité est-elle suffisante et/ou pertinente pour innover dans le domaine des nouvelles technologies ? In G. Valléry & R. Amalberti (Eds.), *L'analyse du travail en perspectives. Influences et évolutions*. Toulouse: Octarès.
Brangier, E., & Bastien, J. M. C. (2010). Ergonomie des produits informatiques : faciliter l'expérience utilisateur en s'appuyant sur les notions d'accessibilité, utilisabilité, émotionnalité et d'influençabilité. In G. Vallery, M. Zouinar, & M.-C. Leport (Eds.), *Ergonomie, conception, de produits et services médiatisés* (pp. 307–328). PUF.
Brangier, E., Dinet, J., & Bastien, J. M. C. (2009). La méthode des staffs d'experts de communautés. Orientation théorique, démarche méthodologique et application pratique. *Document Numérique*, *12*(2), 111–132.
Brangier, E., Dufresne, A., & Hammes-Adelé, S. (2009). Approche symbiotique de la relation humain-technologie : perspectives pour l'ergonomie informatique. *Le Travail Humain*, *72*(4), 333–353. doi:10.3917/th.724.0333
Brangier, E., & Robert, J.-M. (2013). L'innovation par l'ergonomie : Eléments d'ergonomie prospective. *innovatiO*, *1*.
Brannen, J. (1992). *Combining Qualitative and Quantitative Approaches : An Overview*. Aldershotx: Avebury.
Brave, S., & Nass, C. (2009). Emotion in Human-Computer interaction. In A. Sears & J. A. Jacko (Eds.), *Human Computer Interaction : Fondamentals* (pp. 53–68). New York: CRC Press.
Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.
Brennan, R. L. (2013). *Generalizability Theory*. Berlin: Springer Science & Business Media.
Breyfogle, F. (1999). *Implementing Six Sigma : Smarter Solutions Using Statistical Methods*. Hoboken, NJ: John Wiley and Sons.
Briggs, R. O., Adkins, M., Mittleman, D. D., Kruse, J., Miller, S., & Nunamaker Jr., J. F. (1998). Technology Transition Model Derived from Field Investigation of GSS Use Aboard the U.S.S. CORONADO. *Journal of Management Information Systems*, *15*(3), 151–196.

Briggs, R. O., & Reinig, B. (2004). Satisfaction attainment theory as a model for value creation. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*. IEEE. doi:10.1109/HICSS.2004.1265063

Briscoe, B., Odlyzko, A., & Tilly, B. (2006, July). Metcalfe's Law is Wrong. *IEEE Spectrum*.

Broadbent, D. E. (1958). *Perception and Communication*. Oxford: Pergamon Press.

Broadwater, R. L., Haynes, R. R., & Mitry, S. A. (1982). Blood pressure and heart rate measuring watch. Washington, DC:

Brockmyer, J. H., Fox, C. M., Curtiss, K. A., McBroom, E., Burkhart, K. M., & Pidruzny, J. N. (2009). The development of the game engagement questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology*, *45*, 624–634.

Broekens, J., & Brinkman, W.-P. (2013). AffectButton: A method for reliable and valid affective self-report. *International Journal of Human-Computer Studies*, *71*(6), 641–667. doi:10.1016/j.ijhcs.2013.02.003

Brooke, J. (1996). SUS: a "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland (Eds.), *Usability Evaluation in Industry* (pp. 189–194). London: Taylor and Francis.

Brookhuis, K. a. (2008). From ergonauts to infonauts: 50 years of ergonomics research. *Ergonomics*, *51*(1), 55–58. doi:10.1080/00140130701801256

Brown, D. E., James, G. D., Nordloh, L., & Jones, A. A. (2003). Job strain and physiological stress responses in nurses and nurse's aides: predictors of daily blood pressure variability. *Blood Press. Monit.*, *8*(6), 237–242.

Brown, E., & Cairns, P. (2004). A grounded investigation of game immersion. In *Extended abstracts of the 2004 conference on Human factors and computing systems - CHI '04* (p. 1297). New York, New York, USA: ACM Press. doi:10.1145/985921.986048

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*(3), 296–322.

Bruner, J. S., Goodnow, J., & Austin, G. (1956). *A Study of Thinking*. New York: Wiley.

Bruseberg, A., & McDonagh-Philp, D. (2002). Focus groups to support the industrial/product designer: a review based on current literature and designers' feedback. *Applied Ergonomics*, *33*(1), 27–38. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11827136

Bruya, B. (2010). *Effortless attention: A new perspective in the cognitive science of attention and action*. Cambridge, MA: MIT Press.

Bryman, A. (2004). *Social research methods*. Oxford: Oxford University Press.

Bryman, A. (2008). Why Do Researchers Integrate/Combine/Mesh/Blend/Mix/Merge/Fuse Quantitative and Qualitative Research? In M. M. Bergman (Ed.), *Advances in mixed methods research: theories and applications* (pp. 86–100). Los Angeles: SAGE. doi:10.4135/9780857024329

Brynjolfsson, E. (1993). The productivity paradox of information technology. *Communications of the ACM*, *36*(12), 66–77. doi:10.1145/163298.163309

Brynjolfsson, E., & Hitt, L. M. (1998). Beyond the productivity paradox. *Communications of the ACM*, *41*(8), 49–55. doi:10.1145/280324.280332

Bryson, B. (2003). *A Short History of Nearly Everything*. New York: Broadway Books.

Burba, N., Bolas, M., Krum, D. M., & Suma, E. A. (2012). Unobtrusive measurement of subtle nonverbal behaviors with the Microsoft Kinect. In *Proceedings of the 2012 IEEE Virtual Reality (VR '12)* (pp. 1–4). Washington, DC, USA,: IEEE Computer Society. doi:10.1109/VR.2012.6180952

Burmester, M., Mast, M., Jäger, K., & Homans, H. (2010). Valence method for formative evaluation of user experience. In K. Halskov & M. G. G. Petersen (Eds.), *Proceedings of the 8th ACM Conference on Designing Interactive Systems (DIS '10)* (pp. 364–367). ACM. doi:10.1145/1858171.1858239

Buscher, G., Dengel, A., & van Elst, L. (2008). Eye movements as implicit relevance feedback. In *Proceeding of the twenty-sixth annual CHI conference extended abstracts on Human factors in computing systems - CHI '08* (p. 2991). New York, New York, USA: ACM Press. doi:10.1145/1358628.1358796

Bustillo, C. (2007). *Emoscope: An Emotional Usability Tool. Formalization and Application in User-Centered Design*. Universitat Ramon Llull.

Butler, D. (2013). When Google got flu wrong. *Nature*, *494*, 155–156. doi:10.1038/494155a

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (Eds.). (2007). *Handbook of psychophysiology* (3rd ed.). Cambridge: Cambridge University Press.

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2007). Psychophysiological Science: Interdisciplinary Approaches to Classic Questions About the Mind. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (3rd ed., pp. 1–16). Cambridge: Cambridge University Press.

Cacioppo, J. T., Tassinary, L. G., & Bernston, G. G. (2007). *The handbook of psychophysiology*. (john T. (University of C. Cacioppo, L. G. (Texas A. & M. U. Tassinary, & G. G. (Ohio S. U. Bernston, Eds.)*Dreaming* (Third edit.). Cambridge University Press.

Callon, M., Joly, P.-B., & Rip, A. (2010). Reinventing innovation. In M. Arentsen (Ed.), *Governance and Innovation* (pp. 19–32). Cheltenham: Edward Elgar.

Cameron, A. M., Massie, A. B., Alexander, C. E., Stewart, B., Montgomery, R. A., Benavides, N. R., … Segev, D. L. (2013). Social media and organ donor registration: The Facebook effect. *American Journal of Transplantation*, *13*(8), 2059–2065. doi:10.1111/ajt.12312

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. doi:10.1037/h0046016

Card, S. (2009). Information visualization. In A. Sears & J. A. Jacko (Eds.), *Human Computer Interaction: Design Issues, Solutions, and Applications* (pp. 181–215). New York: CRC Press.

Card, S. K., Moran, T. P., & Newell, A. (1980a). Computer text-editing: An information-processing analysis of a routine cognitive skill. *Cognitive Psychology*, *12*(1), 32–74. doi:10.1016/0010-0285(80)90003-1

Card, S. K., Moran, T. P., & Newell, A. (1980b). The keystroke-level model for user performance time with interactive systems. *Communications of the ACM*, *23*(7), 396–410. doi:10.1145/358886.358895

Card, S. K., Moran, T. P., & Newell, A. (1983). *The Psychology of Human Computer Interaction*. Hilglade N.J.: Lawrence Erlbaum.

Cardinet, J., Johnson, S., & Pini, G. (2011). *Applying generalizability theory using EduG* (New York.). Routledge.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory : Applications to educational measurement. *Journal of Educational Measurement*, *13*(2), 119–135.

Carmines, E. G., & Woods, J. (2011). Multimethod Research. In M. S. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The SAGE Encyclopedia of Social Science Research Methods* (pp. 678–681). Thousand Oaks, Calif.: SAGE.

Caron-Fasan, M. L. (1998). Cognition et stratégie d'entreprise : l'exploitation individuelle des informations de veille stratégique. In *Actes de la VIIème Conférence Internationale de l'AIMS*.

Carroll, J. (2000). *Making use: Scenario-based design of human-computer interaction*. Cambridge, Massachusetts: The MIT Press.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge: Cambridge University Press.

Carroll, J. M. (1997). Human-computer interaction: psychology as a science of design. *Annual Review of Psychology*, *48*, 61–83. doi:10.1146/annurev.psych.48.1.61

Carroll, J. M. (2013). Human Computer Interaction - brief intro. In M. Soegaard & R. F. Dam (Eds.), *The Encyclopedia of Human-Computer Interaction, 2nd Ed.* Aarhus, Denmark: The Interaction Design Foundation.

Carter, T. J., & Gilovich, T. D. (2010). The Relative Relativity of Material and Experiential Purchases. *Journal of Personality and Social Psychology*, *98*(1), 146–159.

Carù, A., & Cova, B. (2003). Approche empirique de l'immersion dans l'expérience de consommation : les opérations d'appropriation. *Recherches et Applications En Marketing*, *18*(2), 47–65.

Cattell, R. B. (1988). Psychological theory and scientific method. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 3–20). New York: Plenum Press.

Cavallin, H., Martin, W. M., & Heylighen, A. (2007). How relative absolute can be: SUMI and the impact of the nature of the task in measuring perceived software usability. *AI and Society*, *22*(2), 227–235. doi:10.1007/s00146-007-0127-0

Chadwick-Dias, A., McNulty, M., & Tullis, T. (2003). Web usability and age: How design changes can improve performance. In *Proceedings of the 2003 ACM Conference on Universal Usability*. Vancouver, BC.

Chakaveh, S., & Bogen, M. (2007). Media convergence, an introduction. In *Proceedings of the 12th international conference on Human-computer interaction: intelligent multimodal interaction environments* (pp. 811–814). Retrieved from http://portal.acm.org/citation.cfm?id=1769590.1769682

Chapanis, A. (1996). *Human factors in systems engineering*. New York: Wiley.

Chaparro, B. S. (2008). Usability evaluation of a university portal website. *Usability News*, *10*(2), 1–7.

Chapman, P. (1997). *Models of Engagement : Intrinsically Motivated Interaction with Multimedia Learning Software*. University of Waterloo.

Chen, H., Wigand, R., & Nilan, M. (2000). Exploring web users' optimal flow experiences. *Information Technology & …*, *13*(4), 263–281. Retrieved from http://www.emeraldinsight.com/journals.htm?articleid=883534&show=abstract

Chen, S., & Epps, J. (2014). Using Task-Induced Pupil Diameter and Blink Rate to Infer Cognitive Load. *Human–Computer Interaction*, *29*(4), 390–413. doi:10.1080/07370024.2014.892428

Cherchye, L., Moesen, W., Rogge, N., Van Puyenbroeck, T., Saisana, M., & Saltelli, A. (2008). Creating composite indicators with DEA and robustness analysis : the case of the Technology Achievement Index. *Journal of the Operational Research Society*, *59*(2), 239–251.

Chesbrough, H., Vanhaverbeke, W., & West, J. (2006). *Open innovation: Researching a new paradigm*. Oxford: Oxford University Press.

Chin, J. P., Diehl, V. A., & Norman, L. K. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '88* (pp. 213–218). New York, New York, USA: ACM Press. doi:10.1145/57167.57203
Chonchuir, M. N., & McCarthy, J. (2008). The enchanting potential of technology: a dialogical case study of enchantment and the internet. *Personal and Ubiquitous Computing*, *12*(5), 401–409.
Christensen, C. M. (2013). Disruptive Innovation. In M. Soegaard & R. F. Dam (Eds.), *The Encyclopedia of Human-Computer Interaction, 2nd Ed.* Aarhus, Denmark: The Interaction Design Foundation. Retrieved from https://www.interaction-design.org/encyclopedia/disruptive_innovation.html
Christensen, J. F., Olesen, M. H., & Kjaer, J. S. (2005). The industrial dynamics of open innovation - Evidence from the transformation of consumer electronics. *Research Policy*, *34*(10), 1533– 1549.
Christophersen, T., & Konradt, U. (2011). Reliability, validity, and sensitivity of a single-item measure of online store usability. *International Journal of Human-Computer Studies*, *69*(4), 269–280. doi:10.1016/j.ijhcs.2010.10.005
Chung, J., & Gardner, H. J. (2012). Temporal Presence Variation in Immersive Computer Games. *International Journal of Human-Computer Interaction*, *28*(8), 511–529. doi:10.1080/10447318.2011.627298
Churchill, E. F., & Bardzell, J. (2007). From HCI to Media Experience : Methodological Implications. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...but not as we know it* (pp. 213–214). Swinton, UK, UK: British Computer Society.
Clay, A. (2009). *La branche émotion, un modèle conceptuel pour l'intégration de la reconnaissance multimodale d'émotions dans des applications interactives : application au mouvement et à la danse augmentée (Doctoral dissertation)*. Université de ordeau 1.
Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research*, *18*(1), 115–126.
Clubb, O. L. (2007). Human-to-Computer-to-Human Interactions (HCHI) of the communications revolution. *Interactions*, *14*(2), 35. doi:10.1145/1229863.1229883
Coan, J., & Allen, J. (2007). *Handbook of emotion elicitation and assessment*. Oxford: Oxford University Press.
Cockton, G., & Woolrych, A. (2002). Sale must end: should discount methods be cleared off HCI's shelves? *Interactions*, *9*(5). doi:10.1145/566981.566990
Cody-Allen, E., & Kishore, R. (2006). An extension of the UTAUT model with e-quality, trust, and satisfaction constructs. In *Proceedings of the 2006 ACM SIGMIS CPR Conference on Computer Personnel Research: Forty Four Years of Computer Personnel Research: Achievements, Challenges & the Future* (pp. 82–89). New York, NY: ACM. Retrieved from http://portal.acm.org/citation.cfm?id=1125196
Cohen, B. H., Davidson, R. J., Senulis, J. A., Saron, C. D., & Weisman, D. R. (1992). Muscle tension patterns during auditory attention. *Biological Psychology*, *33*(2-3), 133–156.
Cohen, J. (1960). A coefficient of agreement of nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. doi:10.1177/001316446002000104
Cohen, W. M., & Levinthal, D. A. (1990). Absorptive Capacity: A New Perspective on Learning and Innovation. *Administrative Science Quaterly*, *35*(1), 128–152.
Colin, N., & Verdier, H. (2012). *L'Âge de la multitude : Entreprendre et gouverner après la révolution numérique*. Paris: Armand Colin.
Collin, P., & Colin, N. (2013). *Mission d'expertise sur la fiscalité de l'économie numérique*. Paris, France. Retrieved from http://www.economie.gouv.fr/files/rapport-fiscalite-du-numerique_2013.pdf
Convertino, G., Kannampallil, T. G., & Councill, I. (2006). Mapping the intellectual landscape of CSCW research. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*.
Convertino, G., Neale, D. C., Hobby, L., Carroll, J. M., & Rosson, M. B. (2004). A laboratory method for studying activity awareness. In *Proceedings of the third Nordic conference on Human-computer interaction - NordiCHI '04* (pp. 313–322). New York, New York, USA: ACM Press. doi:10.1145/1028014.1028063
Cook, T. D., & Campbell, D. T. (1979a). *Quasi-experimentation: Design and Analysis for Field Settings*. Boston: Houghton Mifflin.
Cook, T. D., & Campbell, D. T. (1979b). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.
Cooper, A. (1999). *The inmates are running the asylum*. Indianapolis, IN: Sams.
Cooper, R. G., & Kleinschmidt, E. J. (2000). New Product Performance: What Distinguishes the Star Products. *Australian Journal of Management*, *25*(1), 17–46. Retrieved from http://www.questia.com/PM.qst?a=o&se=gglsc&d=5002355633
Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., … Players, F. (2010). Predicting protein structures with a multiplayer online game. *Nature*, *466*(7307), 756–60. doi:10.1038/nature09304
Cordes, R. (1993). The effects of running fewer subjects on time-on-task measures. *International Journal of Human-Computer Interaction*, *5*(4), 393–403.

Cote, J. A., & Buckley, M. R. (1987). Estimating trait, method, and error variance : Generalizing across 70 construct validation studies. *Journal of Marketing Research*, *24*(3), 315–318.
Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & SchroÌder, M. (2000). `FEELTRACE': An Instrument for Recording Perceived Emotion in Real Time. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion* (pp. 19–24). Belfast.
Cowley, B., Charles, D., Black, M., & Hickey, R. (2008). Toward an understanding of flow in video games. *Computers in Entertainment*, *6*(2). doi:10.1145/1371216.1371223
Craine, K. (2000). *Designing a document strategy*. Hurst, Tex: MC2 Books.
Crawford, K. (2013). The Hidden Biases in Big Data. *Havard Business Review*.
Creswell, J. W. (2003). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Thousand Oaks, CA: Sage.
Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.
Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd Edn.). New York: Happer and Row publishers.
Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements : Theory of generalizability for scores and profiles*. New York: Wiley.
Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.
Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of Generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137–163.
Cronbach, L. J., & Shavelson, R. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*, 391–418.
Cronin, B. (1995). Shibboleth and substance in North American Library and Information Science education. *Libri*, *45,* 45–63.
Cruz, A., Correia, A., Paredes, H., Fonseca, B., Morgado, L., & Martins, P. (2012). Towards an overarching classification model of CSCW and groupware: a socio-technical perspective. In V. Herskovic, H. U. Hoppe, M. Jansen, & J. Ziegler (Eds.), *18th international conference on Collaboration and Technology (CRIWG'12)* (pp. 41–56). Berlin, Heidelberg: Springer-Verlag. doi:10.1007/978-3-642-33284-5_4
Csikszentmihalyi, M. (1975). *Beyond Boredom and Anxiety: Experiencing Flow in Work and Play*. New York: Harper Collins.
Csikszentmihalyi, M. (1990). *Flow, the Psychology of Optimal Experience*. New York: Harper and Row.
Csikszentmihalyi, M. (2006). *La créativité: Psychologie de la découverte et de l'invention*. Paris: R. Laffont.
Csikszentmihalyi, M., & Csikszentmihalyi, I. (1988). *Optimal Experience. Psychological studies of Flow in Consciousness*. New York: Cambridge University Press.
Csikszentmihalyi, M., & Larson, R. (1992). Validity and reliability of the experience sampling method. In M. W. DeVries (Ed.), *The Experience of Psychopathology: Investigating Mental Disorders in their Natural Settings* (pp. 43–57). Cambridge, UK: Cambridge University Press.
Cui, W., & Rau, P. L. P. (2012). Comparing the Psychological and Physiological Measurement of Player's Engaging Experience in Computer Game. In *4th AHFE International Conference 2012*.
Cukier, K., & Mayer-Schoenberger, V. (2013). Rise of Big Data: How it's Changing the Way We Think about the World. *Foreign Affairs*, *92*(28).
Czerwinski, M., Horvitz, E., & Cutrell, E. (2001). Subjective Duration Assessment : An Implicit Probe for Software Usability. In *Proceedings of IHM-HCI 2001* (pp. 167–170). Lille, France.
D'Argembeau, A., Ruby, P., Collette, F., Degueldre, C., Balteau, E., Luxen, A., … Salmon, E. (2007). Distinct regions of the medial prefrontal cortex are associated with self-referential processing and perspective taking. *Journal of Cognitive Neuroscience*, *19*(6), 935–944. doi:10.1162/jocn.2007.19.6.935
D'Aveni, R. (1997). Waking up to the new era of hypercompetition. *The Washington Quarterly*, *21*(1), 183–195.
D'Mello, S. K., & Graesser, A. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction*, *20*(2), 147–187. doi:10.1007/s11257-010-9074-4
Daher, L. A., & Elkabani, Ii. (2012). Usability evaluation of some lebanese universities web portals. In *The 13th International Arab Conference on Information Technology ACIT ' 2012*.
Daly, M. (2003). Methodology. In R. L. Miller & J. D. Brewer (Eds.), *The A-Z of Social Research* (pp. 192–194). London: SAGE Publications, Ltd. doi:10.4135/9780857020024
Damodaran, L. (1996). User involvement in the systems design process-a practical guide for users. *Behaviour & Information Technology*, *15*(6), 363–377. doi:10.1080/014492996120049
Darses, F., & Reuzeau, F. (2004). Participation des utilisateurs à la conception des systèmes et dispositifs de travail. In P. Falzon (Ed.), *Ergonomie* (pp. 405–420). Paris: Presses Universitaires de France.
Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319. doi:10.2307/249008

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1992). Extrinsic and Intrinsic Motivation to Use Computers in the Workplace. *Journal of Applied Social Psychology*, *22*, 1111–1132.
Davis, J. P., Steury, K., & Pagulayan, R. A. (2005). Survey method for assessing perceptions of a game: The consumer playtest in game design. *Game Studies*, *5*(1).
Dawson, M. E., Schell, A. M., & Filion, D. L. (2007). The Electrodermal System. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (3rd ed., pp. 159–181). Cambridge: Cambridge University Press.
Dawson, R., & Bynghall, S. (2012). *Getting results from crowds: The definitive guide to using crowdsourcing to grow your business*. Sydney: Advanced Human Technologies.
De Choudhury, M., Gamon, M., & Counts, S. (2012). Happy, nervous or surprised? classification of human affective states in social media. In *Sixth International AAAI Conference on Weblogs and Social Media*.
De Freitas, S., & Griffiths, M. (2008). The convergence of gaming practices with other media forms: what potential for learning? A review of the literature. *Learning, Media and Technology*, *33*(1), 11–20. doi:10.1080/17439880701868796
De Jaucourt, L. (1751). Préjugé. In D. Diderot & J. L. R. D'Alembert (Eds.), *re raisonné des sciences, des arts et des métiers* (p. 13:284). Paris: Le Breton.
De Rosnay, J. (2006). *La Révolte du pronétariat, de (Fayard, 2006)*. Paris: Fayard.
De Sola Pool, I. (1983). *Technologies of freedom*. Cambridge, MA: Harvard University Press.
De Vasconcelos, L. G., & Baldochi, L. A. (2012). Towards an automatic evaluation of web applications. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing - SAC '12* (pp. 709–716). doi:10.1145/2245276.2245410
Dean, D., DiGrande, S., Field, D., Lundmark, A., O'Day, J., Pineda, J., & Zwillenberg, P. (2012). *The Internet Economy in the G-20: The $4.2 Trillion Growth Opportunity*.
DeGusta, M. (2012). Are Smart Phones Spreading Faster than Any Technology in Human History? *MIT Technological Review*.
Dehaene, S. (2012). *Le cerveau statisticien : la révolution bayésienne en sciences cognitives*. Paris, France.
Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Viking Press.
Delbecq, D. (2014). L'illusion tactile, une révolution en marche. *Journal Du CNRS*. Retrieved from https://lejournal.cnrs.fr/articles/lillusion-tactile-une-revolution-en-marche
Demers, D. (2005). *Dictionary of Mass Communication and Media Research: a guide for students, scholars and professionals*. Spokane, WA: Marquette.
Den Uyl, M., van Kuilenburg, H., & Lebert, E. (2005). FaceReader: an online facial expression recognition system. In *th International Conference on Methods and Techniques in Behavioral Research (Measuring Behavior 2005)*. Wageningen, The Netherlands.
Dentsu. (1985). *Kansei shouhi, risei shouhi [Kansei consumption, logic consumption]*. Tokyo, Japan: Nihon Keizai Shinbunsha.
Denzin, N. K. (1978). *The Research Act (2d ed.)*. New York.: McGraw-Hill.
Desai, M., Fukuda-Parr, S., Johansson, C., & Sagasti, F. (2002). Measuring the Technology Achievement of Nations and the Capacity to Participate in the Networking Age. *Journal of Human Development*, *3*(1), 301–311.
Desmet, P., & Hekkert, P. (2007). Framework of product experience. *International Journal of Design*, *1*(1), 57–66. Retrieved from http://www.ijdesign.org/ojs/index.php/IJDesign/article/view/66
Desmet, P., Overbeeke, K., & Tax, S. (2001). Designing products with added emotional value: development and application of an approach for research through design. *The Design Journal*, *4*(1), 32–47. doi:10.2752/146069201789378496
Desurvire, H., Caplan, M., & Toth, J. A. (2004). Using heuristics to evaluate the playability of games. In *Extended abstracts of the 2004 conference on Human factors and computing systems - CHI '04* (p. 1509). New York, New York, USA: ACM Press. doi:10.1145/985921.986102
Desurvire, H. W., Kondziela, J. M., & Atwood, M. E. (1993). What is gained and lost when using evaluation methods other than empirical testing. In A. Monk, D. Diaper, & M. D. Harrison (Eds.), *Proceedings of the conference on People and computers VII (HCI'92)* (pp. 89–102). New York, NY, USA: Cambridge University Press.
Desurvire, H., & Wiberg, C. (2009). Game Usability Heuristics (PLAY) for Evaluating and Designing Better Games : The Next Iteration. In A. A. Ozok & P. Zaphiris (Eds.), *Proceedings of the 3d International Conference on Online Communities and Social Computing: Held as Part of HCI International 2009 (OCSC '09)* (pp. 557–566). Berlin, Heidelberg: Springer-Verlag.
Dickie, G. (1988). *Evaluating Art*. Philadelphia: Temple University Press.
Diefenbach, S., & Hassenzahl, M. (2009). The "Beauty Dilemma": Beauty is Valued but Discounted in Product Choice. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI '09* (p. 1419). New York, New York, USA: ACM Press. doi:10.1145/1518701.1518916

Dimberg, U. (1990). Facial electromyography and emotional reactions. *Psychophysiology*, *27*(5), 481–494. doi:10.1111/j.1469-8986.1990.tb01962.x
Dirican, A. C., & Göktürk, M. (2012). Involuntary postural responses of users as input to Attentive Computing Systems: An investigation on head movements. *Computers in Human Behavior*, *28*(5), 1634–1647. doi:10.1016/j.chb.2012.04.002
Donnelly, R., & Gardner, J. (2011). Content analysis of computer conferencing transcripts. *Interactive Learning Environments*, *19*(4), 303–315. doi:10.1080/10494820903075722
Dorfman, D. D., & McKenna, H. (1966). Pattern preference as a function of pattern uncertainty. *Canadian Journal of Psychology*, *20*, 143–153.
Doya, K. (2007). *Bayesian brain: Probabilistic approaches to neural coding*. Cambridge, Mass: MIT Press.
Dretske, F. (2008). Epistemology and Information. In P. Adriaans & J. van Benthem (Eds.), *Philosophy of information*. UK: North-Holland.
Drucker, P. F. (1959). *Landmarks of Tomorrow*. New York: Harper & Brothers.
Dumas, J. S., & Fox, J. E. (2009). Usability testing: current practice and future directions. In A. Sears & J. A. Jacko (Eds.), *Human Computer Interaction : Development Process* (pp. 231–251). New York: CRC Press.
Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10937327
Eid, M., & Diener, E. (2006). Introduction: the need for multimethod measurement in psychology. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology*. Washington, DC: American Psychological Association.
Eid, M., & Dienier, E. (2001). Norm for experiencing emotion in different cultures. *Journal of Personality and Social Psychology*, *81*(5), 869–885.
Eiglier, P. (2004). *Marketing et stratégie des services*. Paris: Economica.
Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3/4), 169–200.
Ekman, P., Davidson, R. J., & Friesen, W. V. (1990). The Duchenne smile: Emotional expression and brain physiology II. *Journal of Personality and Social Psychology*, *58*, 342–353.
Ekman, P., Levenson, R. W., & Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science*, *221*(4616), 1208–1210.
Ellis, C. A., Gibbs, S. J., & Rein, G. (1991). Groupware: some issues and experiences. *Communications of the ACM*, *34*(1), 39–58. doi:10.1145/99977.99987
Embretson, S., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc. Publishers.
ENGAGE. (2006). *Report on the evaluation of generative tools and methods for "emotional design". Deliverable D15.3.*
Engelbart, D. C. (1962). *Augmenting human intellect : a conceptual framework. Summary report AFOSR-3233*.
Engeser, S., & Rheinberg, F. (2008). Flow, performance and moderators of challenge-skill balance. *Motivation and Emotion*, *32*(3), 158–172. doi:10.1007/s11031-008-9102-4
English, J. (2005). *The Economy of Prestige*. Cambridge, Mass.: Harvard University Press.
English, W. K., Engelbart, D. C., & Berman, M. L. (1967). Display-Selection Techniques for Text Manipulation. *IEEE Transactions on Human Factors in Electronics*, *HFE-8*(1), 5–15. doi:10.1109/THFE.1967.232994
Epstein, S. (1986). Does aggregation produce spuriously high estimates of behavior stability ? *Journal of Personality and Social Psychology*, *50*(6), 1199–1210.
Ereback, A.-L., & Höök, K. (1994). Using cognitive walkthrough for evaluating a CSCW application. In *Conference companion on Human factors in computing systems - CHI '94* (pp. 91–92). New York, New York, USA: ACM Press. doi:10.1145/259963.260065
Erevelles, S. (1998). The role of affect in marketing. *Journal of Business Research*, *42*(3), 199–215.
Ermi, L., & Mäyrä, F. (2005). Fundamental Components of the Gameplay Experience : Analysing Immersion. In *Proceedings of DiGRA 2005 Conference: Changing Views – Worlds in Play*.
Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.
Essa, I. (1999). Computers seeing people. *AI Magazine*, *20*(2), 69–82.
Express Roularta Services. (2011). *Without Information are we nothing ? Etude sur les tendances de consommation des Français pour l'Express et Illigo*.
Fabernovel. (2014). *GAFAnomics : New Economy, New Rules*. Retrieved from http://www.fabernovel.com/fr/gafa/
Facebook. (2009). *Company timeline*. Retrieved from http://www.facebook.com/press/ info.php?timeline
Facebook. (2013). *Facebook Reports First Quarter 2013 Results*. Retrieved from http://investor.fb.com/releasedetail.cfm?ReleaseID=761090

Fairclough, S. H. (2009). Fundamentals of physiological computing. *Interacting with Computers*, *21*(1-2), 133–145. doi:10.1016/j.intcom.2008.10.011
Falzon, P. (2005). Ergonomie, conception et développement. In *Conférence introductive, 40ème Congrès de la SELF*.
Fan, J., McCandliss, B. D., Fossella, J., Flombaum, J. I., & Posner, M. I. (2005). The activation of attentional networks. *Neuroimage*, *26*, 471–479.
Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience*, *14*(3), 340–347.
Fang, Y. M., & Sun, M. H. (2014). The Exploratory Study of Emotional Valence and Arousal for Eco-visualization Interface of Water Resources. In *Communications in Computer and Information Science* (Vol. 434 PART I, pp. 311–316). Springer International Publishing. doi:10.1007/978-3-319-07857-1_55
Federoff, M. (2002). *Heuristics and Usability Guidelines for the Creation and Evaluation of Fun in Video Games*. Indiana University Master of Science, Bloomington, IN.
Fenouillet, F. (2012). *Les théories de la motivation*. Paris: Dunod.
Filippi, S., & Barattin, D. (2012). Generation, Adoption, and Tuning of Usability Evaluation Multimethods. *International Journal of Human-Computer Interaction*, *28*(6), 406–422. doi:10.1080/10447318.2011.607421
Finstad, K. (2010). The Usability Metric for User Experience. *Interacting with Computers*, *22*(5), 323–327. doi:10.1016/j.intcom.2010.04.004
Finstad, K. (2013). Response to commentaries on "The Usability Metric for User Experience." *Interacting with Computers*, *25*, 327–330.
Firat, A. F., & Venkatesh, A. (1995). Liberatory Postmodernism and the Reenchantment of Consumption. *Journal of Consumer Research*, *22*(3), 239–267.
Foelstad, A. (2007). Group-based Expert Walkthrough. In D. Scapin & E. L. Law (Eds.), *R³UEMs: Review, Report and Refine Usability Evaluation Methods* (pp. 58–60).
Fogg, B. J. (2003). *Persuasive Technology: Using Computers to Change What We Think and Do*. Boston, MA: Morgan Kaufmann.
Fogg, B. J. (2008). Mass Interpersonal Persuasion: An Early View of a New Phenomenon. In H. Oinas-Kukkonen, P. Hasle, M. Harjumaa, K. Segerståhl, & P. Øhrstrøm (Eds.), *Proceedings of the 3rd international conference on Persuasive Technology (PERSUASIVE '08)* (Vol. 5033, pp. 23–34). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-68504-3
Forlizzi, J., & Battarbee, K. (2004). Understanding experience in interactive systems. In *Proceedings of the 2004 conference on Designing interactive systems processes, practices, methods, and techniques - DIS '04* (pp. 261–268). New York, New York, USA: ACM Press. doi:10.1145/1013115.1013152
Fornerino, M., Helme-Guizon, A., & Gotteland, D. (2006). Mesurer l'immersion dans une expérience de consommation : Premiers développements. In *XXIIème Congrès de l'Association Française du Marketing*. Nancy.
Fox Keller, E. (2000). *The Century of the Gene*. Cambridge, MA: Harvard University Press.
Fr0kjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '00* (pp. 345–352). New York, NY, USA: ACM. doi:10.1145/332040.332455
Frackowiak, R., Friston, K., Frith, C., Dolan, R., Price, C., & Zeki, S. (2003). Brain systems mediating reward. In *Human Brain Function* (pp. 445–470). San Diego: Academic Press.
Frank, R. H., & Cook, P. J. (1995). *The winner-take-all society*. New York: The Free Press.
Fréchet, M. (2002). *Les conflits dans les partenariats d'innovation*. Toulouse 1.
Freeman, F. G., Mikulka, P. J., Prinzel, L. J., & Scerbo, M. W. (1999). Evaluation of an adaptive automation system using three EEG indices with a visual tracking task. *Biological Psychology*, *50*(1), 61–76. doi:10.1016/S0301-0511(99)00002-2
Frey, L. R., Botan, C. H., Friedman, P. G., & Kreps, G. L. (1992). *Interpreting Communication Research: A Case Study Approach*. London, UK: Pearson.
Frijda, N. H. (1986). *The emotions*. Cambridge, UK: Cambridge University Press.
Frischen, A., Bayliss, A. P., & Tipper, S. P. (2007). Gaze Cueing of Attention : Visual Attention , Social Cognition , and Individual Differences. *Psychological Bulletin*, *133*(4), 694 –724. doi:10.1037/0033-2909.133.4.694
Fritsche, I., & Linneweber, V. (2006). Nonreactive methods in psychological research. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology*. Washington, DC: American Psychological Association.
Fritz, C., Curtin, J., Poitevineau, J., Morrel-Samuels, P., & Tao, F.-C. (2012). Player preferences among new and old violins. *Proceedings of the National Academy of Sciences*, *109*(3), 760–763. doi:10.1073/pnas.1114999109

Frodon, J.-M. (1998). *La projection nationale : cinéma et nation*. Paris, France: Editions Odile Jacob.
Froehlich, J., Chen, M. Y., Consolvo, S., Harrison, B., & Landay, J. A. (2007). MyExperience : a system for in situ tracing and capturing of user feedback on mobile phones. In *Proceedings of the 5th international conference on Mobile systems, applications and services* (Vol. San Juan, pp. 57–70). New York: ACM Press. doi:10.1145/1247660.1247670
Fu, F., Su, R., & Yu, S. (2009). EGameFlow: A scale to measure learners' enjoyment of e-learning games. *Computers & Education*, *52*(1), 101–112. doi:10.1016/j.compedu.2008.07.004
Fujioka, W. (1984). *Sayonara taishuu kansei jidai wo dou yomu ka [Goodbye, mass – How to read Kansei age?]*. Kyoto, Japan: PHP Research Center.
Fung, K. (2013). *Numbersense : How to use big data to your advantage*. New York: McGraw-Hill Education.
Gaines, B. R. (1985). From Ergonomics to the Fifth Generation: 30 Years of Human-Computer Interaction. *Computer Compacts*, *2*(5-6), 158–161.
Galton, F. (1907). Vox populi. *Nature*, *75*, 450–451.
Ganglbauer, E., Schrammel, J., Deutsch, S., & Tscheligi, M. (2011). Applying Psychophysiological Methods for Measuring User Experience: Possibilities, Challenges and Feasibility. In P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, & M. Winckler (Eds.), *Human-Computer Interaction – INTERACT 2011* (Vol. 6949). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-23768-3
Gao, Y., Barreto, A., & Adjouadi, M. (2009). Monitoring and Processing of the Pupil Diameter Signal for Affective Assessment of a Computer User. In J. A. Jacko (Ed.), *Human-Computer Interaction. New Trends* (Vol. 5610, pp. 49–58). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-02574-7
Gautier, F., & Lenfle, S. (2004). L'avant-projet : définitions et enjeux. In G. Garel, V. Giard, C. Midler, & R. Calvi (Eds.), *Faire de la recherche en management de projet* (pp. 11–33). Paris: Vuibert.
Gaver, B., & Martin, H. (2000). Alternatives : Exploring Information Appliances through Conceptual Design Proposals. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '00* (pp. 209–216). New York, New York, USA: ACM Press. doi:10.1145/332040.332433
Gegner, L., & Runonen, M. (2012). For What it is Worth: Anticipated eXperience Evaluation. In *8th International Conference on Design and Emotion*. London, UK.
Gershon, N. (1995). Human Information Interaction. In *WWW4 Conference*.
Ghani, J. A., & Deshpande, S. P. (1994). Task Characteristics and the Experience of Optimal Flow in Human-Computer Interaction. *The Journal of Psychology*, *128*(4), 381–391.
Ghani, J. A., Supnick, R., & Rooney, P. (1991). The Experience of Flow in Computer-Mediated and in Face-to-Face Groups. In J. I. DeGross, I. Benbasat, G. DeSanctis, & C. M. Beath (Eds.), *Proceedings of the Twelfth International Conference on Information Systems* (pp. 229–238). New York.
Giles, G. (2005). Internet encyclopaedias go head to head. *Nature*, *438*, 900–901.
Gillier, T. (2010). *Comprendre la génération des objets de coopération interentreprises par une théorie des co-raisonnements de conception: vers une nouvelle ingénierie des partenariats d'exploration technologique*. Institut National Polytechnique de Lorraine-INPL.
Gillier, T., & Piat, G. (2008). Co-Designing Broad Scope of Technology-Based Applications in an Exploratory Partnership. In *DS 48: Proceedings DESIGN 2008, the 10th International Design Conference*. Dubrovnik.
Gillier, T., & Piat, G. (2011). Exploring Over: The Presumed Identity of Emerging Technology. *Creativity and Innovation Management*, *20*(4), 238–252. doi:10.1111/j.1467-8691.2011.00614.x
Gillmor, D., & Noren, A. (2004). *We the Media: Grassroots Journalism by the People, for the People*. Sebastopol, CA: O'Reilly Media, Inc.
Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings of instruction : estimation of the teacher and course components. *Journal of Educational Measurement*, *15*(1), 1–13.
Gilroy, S. W., Cavazza, M., & Vervondel, V. (2011). Evaluating Multimodal Affective Fusion using Physiological Signals. In *Proceedings of the 16th international conference on Intelligent user interfaces - IUI '11* (pp. 53–62). New York, New York, USA: ACM Press.
Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*, 1012–1014. doi:10.1038/nature07634
Ginsburgh, V. (2003). Awards, Success and Aesthetic Quality in the Arts. *Journal of Economic Perspectives*, *17*(2), 99–111. doi:10.1257/089533003765888458
Ginsburgh, V., & Weyers, S. (1999). On the perceived quality of movies. *Journal of Cultural Economics*, *23*(4), 269–283.
Glass, G. V, & Ellett, F. S. (1980). Evaluation Research. *Annual Review of Psychology*, *31*(1), 211–228. doi:10.1146/annurev.ps.31.020180.001235
Gleick, J. (2012). *The Information: A History, a Theory, a Flood*. Vintage.
Gombrich, E. (1966). *Norm and Form. Studies in the Art of the Renaissance*. London: Phaidon.

Gonguet, A., Martinot, O., Rodio, F., & Hiribarren, V. (2013). SlideWorld: A Multidisciplinary Research Project to Reinvent the Videoconferencing User Experience. *Bell Labs Technical Journal*, *17*(4), 133–144. doi:10.1002/bltj.21579

Gonguet, A., & Rodio, F. (2011). Problèmes soulevés par la capture et la représentation des émotions pour les systèmes de communication immersifs. In *Companion Proceedings of IHM'11, 23ème Conférence francophone sur les Interactions Homme-Machine*. Sophia Antipolis, France.

González, M. P., Lorés, J., & Granollers, A. (2008). Enhancing usability testing through datamining techniques: A novel approach to detecting usability problem patterns for a context of use. *Information and Software Technology*, *50*(6), 547–568. doi:10.1016/j.infsof.2007.06.001

Goodman, D. (1987). *The complete HyperCard handbook*. New York, NY: Bantam Books.

Grassi, M., Cambria, E., Hussain, A., & Piazza, F. (2011). Sentic Web: A New Paradigm for Managing Social Media Affective Information. *Cognitive Computation*, *3*(3), 480–489. doi:10.1007/s12559-011-9101-8

Gray, W. D., John, B. E., & Atwood, M. E. (1992). The precis of Project Ernestine or an overview of a validation of GOMS. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '92* (pp. 307–312). New York, New York, USA: ACM Press. doi:10.1145/142750.142821

Gray, W. D., John, B. E., Stuart, R., Lawrence, D., & Atwood, M. E. (1990). GOMS meets the phone company: Analytic modeling applied to real-world problems. In D. Diaper, D. J. Gilmore, G. Cockton, & B. Shackel (Eds.), *In Proceedings of the IFIP TC13 Third Interational Conference on Human-Computer Interaction (INTERACT '90)* (pp. 29–34). Amsterdam, The Netherlands: North-Holland Publishing Co.

Gray, W., & Salzman, M. (1998). Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods. *Human-Computer Interaction*, *13*(3), 203–261. doi:10.1207/s15327051hci1303_2

Green, W. S., & Jordan, P. W. (1999). *Human factors in product design, current practice and future trends*. London: Taylor & Francis.

Greenberg, S., & Buxton, B. (2008). Usability evaluation considered harmful (some of the time). In *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08* (p. 111). New York, New York, USA: ACM Press. doi:10.1145/1357054.1357074

Greene, J. C. (2005). Mixed Methods. In S. Mathison (Ed.), *Encyclopedia of Evaluation*. Thousand Oaks, Calif.: Sage. doi:10.4135/9781412950558

Greenwald, a. G. (2012). There Is Nothing So Theoretical as a Good Method. *Perspectives on Psychological Science*, *7*(2), 99–108. doi:10.1177/1745691611434210

Gregory, Rc. I. d'Études et de R. sur l'Expressichard. (1983). De la représentation dans ses rapports avec les illusions sensorielles. In L. Roux (Ed.), *Dénominateurs communs aux arts et aux sciences*. Saint-Étienne: Centre Interdisciplinaire d'Études et de Recherche sur l'Expression contemporaine.

Grewe, O., Nagel, F., Kopiez, R., & Altenmu, E. (2007). Emotions Over Time : Synchronicity and Development of Subjective , Physiological , and Facial Affective Reactions to Music. *Emotion*, *7*(4), 774 –788. doi:10.1037/1528-3542.7.4.774

Gross, M. M., Crane, E. a., & Fredrickson, B. L. (2010). Methodology for Assessing Bodily Expression of Emotion. *Journal of Nonverbal Behavior*, *34*(4), 223–248. doi:10.1007/s10919-010-0094-x

Grudin, J. (2005). Three Faces of Human-Computer Interaction. *IEEE Ann. Hist. Comput.*, *27*(4), 46–62.

Grudin, J. (2012a). A Moving Target — The Evolution of Human-Computer Interaction. In J. Jacko (Ed.), *Human-computer interaction handbook: Fundamentals, evolving technologies, and emerging applications* (3rd editio., pp. 1–40). CRC Press.

Grudin, J. (2012b). Punctuated equilibrium and technology change. *Interactions*, *19*(5), 62. doi:10.1145/2334184.2334200

Grudin, J., & Poltrock, S. (1995). Software engineering and the CHI & CSCW communities. In R. Taylor & J. Coutaz (Eds.), *Software Engineering and Human-Computer Interaction SE - 10* (Vol. 896, pp. 93–112). Springer Berlin Heidelberg. doi:10.1007/BFb0035809

Grudin, J., & Poltrock, S. (2012). Taxonomy and Theory in Computer Supported Cooperative Work. In S.W. Kozlowski (Ed.), *Handbook of organizational psychology* (pp. 1323–1348). Oxford University Press.

Grunert, K. G., & Bech-Larsen, T. (2005). Explaining choice option attractiveness by beliefs elicited by the laddering method. *Journal of Economic Psychology*, *26*(2), 223–241. doi:10.1016/j.joep.2004.04.002

Guerlesquin, G. (2012). *Articulation Ergonomie-Design-Conception Mécanique : approche méthodologique de la convergence multidisciplinaire*. Université de Technologie de Belfort-Montbeliard.

Guilford, J. P. (1967). *The nature of human intelligence*. New York: McGraw-Hill.

Gustafsson, J. E., & Snow, R. E. (1997). Ability profiles. In R. F. Dillon (Ed.), *Handbook on testing* (pp. 107–135). Westport, CT, US: Greenwood Press/Greenwood Publishing Group.

Gutman, J. (1982). A means-end chain model based on consumer categorization processes. *Journal of Marketing*, *46*(2), 60–72.

Gutwin, C., & Greenberg, S. (1999). The effects of workspace awareness support on the usability of real-time distributed groupware. *ACM Transactions on Computer-Human Interaction*, *6*(3), 243–281. doi:10.1145/329693.329696

Gutwin, C., & Greenberg, S. (2000). The Mechanics of Collaboration: Developing Low Cost Usability Evaluation Methods for Shared Workspaces. In *Proceedings of the 9th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE '00)* (pp. 98–103). Washington, DC, USA: IEEE Computer Society.

Gutwin, C., Roseman, M., & Greenberg, S. (1996). A usability study of awareness widgets in a shared workspace groupware system. In *Proceedings of the 1996 ACM conference on Computer supported cooperative work - CSCW '96* (pp. 258–267). New York, New York, USA: ACM Press. doi:10.1145/240080.240298

Haapalainen, E., Kim, S., Forlizzi, J. F., & Dey, A. K. (2010). Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM international conference on Ubiquitous computing - Ubicomp '10* (p. 301). New York, New York, USA: ACM Press. doi:10.1145/1864349.1864395

Hackos, J. T., & Redish, J. C. (1988). *User and task analysis for interface design*. New York, NY: John Wiley & Sons, Inc.

Hammer, M. (1984). The OA mirage. *Datamation*, *30*(2), 36–46.

Hancock, P. A., & Warm, J. S. (1989). A dynamic model of stress and sustained attention. *Human Factors*, *31*, 519–537.

Hancock, P. a., & Weaver, J. L. (2005). On time distortion under stress. *Theoretical Issues in Ergonomics Science*, *6*(2), 193–211. doi:10.1080/14639220512331325747

Hand, D. J. (2004). *Measurement theory and practice: The world through quantification*. London: Arnold.

Hansen, A. L., Johnsen, B. H., & Thayer, J. F. (2003). Vagal influence on working memory and attention. *International Journal of Psychophysiology*, *48*(3), 263–274.

Harford, T. (2014). Big data: are we making a big mistake? *Financial Times*, pp. 7–11. doi:10.1111/j.1740-9713.2014.00778.x

Hargadon, A., & Sutton, R. I. (2001). Building an innovation factory. *Harvard Business Review*, *78*(3), 157–166.

Harrison, S., Tatar, D., & Sengers, P. (2007). The Three Paradigms of HCI. In *Alt. CHI, proceedings of the 2007 annual conference on Human factors in computing systems - CHI'07*. doi:10.1234/12345678

Hassenzahl, M. (2003). The thing and I: understanding the relationship between user and product. In M. Blythe, C. Overbeeke, A. F. Monk, & P. C. Wright (Eds.), *Funology: From Usability to Enjoyment* (pp. 31–42). Dordrecht: Kluwer.

Hassenzahl, M. (2004a). Emotions can be quite ephemeral. We cannot design them. *Interactions*, *11*(5), 46–48. doi:10.1145/1015530.1015551

Hassenzahl, M. (2004b). The Interplay of Beauty, Goodness, and Usability in Interactive Products. *Human-Computer Interaction*, *19*(4), 319–349. doi:10.1207/s15327051hci1904_2

Hassenzahl, M. (2006). Hedonic, emotional, and experiential perspectives on product quality. In C. Ghaoui (Ed.) (Ed.), *Encyclopedia of Human Computer Interaction* (pp. 266–272). Idea Group Publishing.

Hassenzahl, M. (2008). User experience (UX): towards an experiential perspective on product quality. In *Proceedings of the 20th International Conference of the Association Francophone d'Interaction Homme-Machine - IHM '08* (pp. 11–15). New York, New York, USA: ACM Press. doi:10.1145/1512714.1512717

Hassenzahl, M. (2013). User Experience and Experience Design. In M. Soegaard & R. F. Dam (Eds.), *The Encyclopedia of Human-Computer Interaction, 2nd Ed.* (2nd Ed.). Aarhus, Denmark: The Interaction Design Foundation.

Hassenzahl, M., Beu, A., & Burmester, M. (2001). Engineering joy. *IEEE Software*, *18*(1), 70–76. doi:10.1109/52.903170

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualitat [AttracDiff: A questionnaire to measure perceived hedonic and pragmatic quality]. In J. Ziegler & G. Szwillus (Eds.), *Mensch & Computer 2003 : Interaktion in Bewegung* (pp. 187–196). Stuttgart, Leipzig: B. G. Teubner.

Hassenzahl, M., Platz, A., Burmester, M., & Lehner, K. (2000). Hedonic and ergonomic quality aspects determine a software's appeal. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '00* (pp. 201–208). New York, USA: ACM Press. doi:10.1145/332040.332432

Hassenzahl, M., & Sandweg, N. (2004). From mental effort to perceived usability: transforming experiences into summary assessments. In *CHI'04 extended abstracts on Human factors in computing systems* (pp. 1283–1286). ACM. Retrieved from http://portal.acm.org/citation.cfm?id=986044

Hassenzahl, M., & Tractinsky, N. (2006). User experience - a research agenda. *Behaviour & Information Technology*, *25*(2), 91–97. doi:10.1080/01449290500330331

Hassenzahl, M., & Wessler, R. (2000). Capturing design space from a user perspective: The repertory grid technique revisited. *International Journal of Human-Computer Interaction*, *12*(3-4), 441–459. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/10447318.2000.9669070

Hatchuel, A., & Weil, B. (2003). A new approach of innovative Design : an introduction to CK theory. In *Proceedings of ICED 03, the 14th International Conference on Engineering Design*. Stockholm.

Haynes, S. R., Purao, S., & Skattebo, A. L. (2004). Situating evaluation in scenarios of use. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work - CSCW '04* (p. 92). New York, New York, USA: ACM Press. doi:10.1145/1031607.1031624

Hazlett, R., & Benedek, J. (2007). Measuring emotional valence to understand the user's experience of software. *International Journal of Human-Computer Studies*, *65*(4), 306–314. doi:10.1016/j.ijhcs.2006.11.005

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience sampling method: measuring the quality of everyday life*. Thousand Oaks, Calif.: Sage Publications. doi:10.4135/9781412984201

Helander, M. (2003). Forget about ergonomics in chair design? Focus on aesthetics and comfort! *Ergonomics*, *46*(13-14), 1306–1319. doi:10.1080/00140130310001610847

Helander, M. G., & Zhang, L. (1997). Field studies of comfort and discomfort in sitting. *Ergonomics*, *40*(9), 895–915. doi:10.1080/001401397187739

Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T., & Weitzner, D. (2008). Web Science: An interdisciplinary approach to understanding the Web. *Communications of ACM*, *51*(7), 62–69.

Herlocker, J. L., Konstan, J. a, & Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work (CSCW '00)* (pp. 241–250). New York, NY, USA: ACM. doi:10.1145/358916.358995

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, *22*(1), 5–53. doi:10.1145/963770.963772

Hertzum, M., & Holmegaard, K. D. (2013). Perceived Time as a Measure of Mental Workload: Effects of Time Constraints and Task Success. *International Journal of Human-Computer Interaction*, *29*(1), 26–39. doi:10.1080/10447318.2012.676538

Hertzum, M., & Jacobsen, N. E. (2003). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, *15*(1), 183–204.

Hertzum, M., Molich, R., & Jacobsen, N. E. (2014). What you get is what you see: revisiting the evaluator effect in usability tests. *Behaviour & Information Technology*, *33*(2), 144–162. doi:10.1080/0144929X.2013.783114

Herzberg, F., Mausner, B., & Snyderman, B. B. (1959). *The Motivation to Work*. New York: John Wiley.

Hess, E., H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science*, *132*, 349–350.

Hetherington, J. (2000). Role of Theory and Experimental Design in Multivariate Analysis and Mathematical Modeling. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of Applied Multivariate Statistics and Mathematical Modeling* (pp. 37–63). San Diego, CA: Elsevier B.V. doi:10.1016/B978-012691360-6/50003

Hetzel, P., & Volle, P. (2002). L'expérientiel : de la théorie à l'action. *Décisions Marketing*, *28*, 5–6.

Higgins, E. T., & Scholer, A. a. (2009). Engaging the consumer: The science and art of the value creation process. *Journal of Consumer Psychology*, *19*(2), 100–114. doi:10.1016/j.jcps.2009.02.002

Hine, T. (1995). *The total package: The evolution and secret meanings of boxes, bottles, cans, and tubes*. Boston: Little Brown.

Hirschman, E., & Pieros, A. (1985). Relationships among indicators of success in Broadway plays and motion pictures. *Journal of Cultural Economics*, *9*(1), 35–63.

Hodgson, R. (2008). An examination of judge reliability at a major US wine competition. *Journal of Wine Economics 3, 3*(2), 105–113.

Hogan, J., & Roberts, B. W. (1996). Issues and non-issues in the fidelity–bandwidth trade-off. *Journal of Organizational Behavior*, *17*(6), 627–637. doi:10.1002/(SICI)1099-1379(199611)17:6<627::AID-JOB2828>3.0.CO;2-F

Holbrook, M. B., & Hirschman, E. C. (1982). The Experiential Aspects of Consumption: Consumer Fantasies, Feelings, and Fun. *Journal of Consumer Research*, *9*(2), 132. doi:10.1086/208906

Holiday, R. (2014). *Growth hacker marketing: A primer on the future of PR, marketing, and advertising*. New York: Portfolio/Penguin.

Holt, D. (1995). How consumers consume : a typology of consumption practices. *Journal of Consumer Research*, *22*(1), 1–16.

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229–247. doi:10.1111/j.1540-5907.2009.00428.x

Hopper, G. M. (1952). The education of a computer. In *Proceedings of the 1952 ACM national meeting (Pittsburgh) on - ACM '52* (pp. 243–249). New York, New York, USA: ACM Press. doi:10.1145/609784.609818

Hornbæk, K. (2010). Dogmas in the assessment of usability evaluation methods. *Behaviour & Information Technology*, *29*(1), 97–111. doi:10.1080/01449290801939400

Hornbæk, K., & Law, E. L. (2007). Meta-analysis of correlations among usability measures. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07* (p. 617). New York, New York, USA: ACM Press. doi:10.1145/1240624.1240722

Hornbak, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, *64*(2), 79–102. doi:10.1016/j.ijhcs.2005.06.002

Howe, J. (2006, June). The Rise of Crowdsourcing. *Wired*.

Hubbard, E., & Ramachandran, V. S. (2003). The phenomenology of synaesthesia. *Journal of Consciousness Studies*, *10*(8), 49–57.

Hughes, J., King, V., Rodden, T., & Andersen, H. (1994). Moving out from the control room: ethnography in system design. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work - CSCW '94* (pp. 429–439). New York, New York, USA: ACM Press. doi:10.1145/192844.193065

Hughes, J., Sharrock, W., Rodden, T., O'Brien, J., Rouncefield, M., & Calvey, D. (1994). *Field Studies and CSCW*. Lancaster, UK: Lancaster University.

Huizingh, E. K. R. E. (2011). Open innovation: State of the art and future perspectives. *Technovation*, *31*(1), 2–9. doi:10.1016/j.technovation.2010.10.002

Hulin, C., Cudeck, R., Netemeyer, R., Dillon, W. R., McDonald, R., & Bearden, W. (2001). Measurement. *Journal of Consumer Psychology*, *10*(1-2), 55–69. doi:10.1207/S15327663JCP1001&2_05

HUMAINE. (2008a). *D4h: Final report on WP4*. Retrieved from http://emotion-research.net/projects/humaine/deliverables/D4h-Final report on WP4.pdf

HUMAINE. (2008b). *D9j: Final report on WP9*. Retrieved from http://emotion-research.net/projects/humaine/deliverables/D9j-Final report on WP9.pdf

Humphries, W. D., Neale, D. C., McCrickard, D. S., & Carroll, J. M. (2004). Laboratory Simulation Methods for Studying Complex Collaborative Tasks. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*(21), 2451–2455. doi:10.1177/154193120404802102

Huron, S., Isenberg, P., & Fekete, J. D. (2013). PolemicTweet: Video Annotation and Analysis through Tagged Tweets. In P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, & M. Winckler (Eds.), *Human-Computer Interaction – INTERACT 2013* (Vol. 8118, pp. 135–152). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-40480-1_9

Huxley, R. (2010). Music is Not Our Currency. *Creative Deconstruction*. Retrieved from http://www.creativedeconstruction.com/2010/05/music-is-not-our-currency

Hwang, W., & Salvendy, G. (2007). What makes evaluators to find more usability problems? : a meta-analysis for individual detection rates. In J. A. Jacko (Ed.), *Proceedings of the 12th international conference on Human-computer interaction: interaction design and usability (HCI'07)* (pp. 499–507). Berlin, Heidelberg: Springer-Verlag.

IJsselsteijn, W., de Ridder, H., Hamberg, R., Bouwhuis, D., & Freeman, J. (1998). Perceived depth and the feeling of presence in 3DTV. *Displays*, *18*, 207–214.

IJsselsteijn, W., Poels, K., & de Kort, Y. A. W. (2008). *The Game Experience Questionnaire: Development of a self-report measure to assess player experiences of digital games*. Eindhoven: TU Eindhoven. FUGA Deliverable D3.3. Technical report.

Ijsselsteijn, W., Ridder, H. d., Freeman, J., & Avons, S. E. (2000). Presence : Concept, determinants and measurement. In *Proceedings of the SPIE, Human Vision and Electronic Imaging* (pp. 3959–3976).

Ijsselsteijn, W., & Riva, G. (2003). Being There : The experience of presence in mediated environments. In G. Riva, F. Davide, & W. IJsselsteijn (Eds.), *Being there: Concepts, effects and measurements of user presence in synthetic environments*. Amsterdam: IOS Press.

Inglehart, R. (1997). *Modernization and Postmodernization: Cultural, Economic, and Political Change in 43 Societies*. Princeton University Press.

Inkpen, K., Mandryk, R., Dimicco, J., & Scott, S. (2004). Methodology for Evaluating Collaboration Behaviour in Co-Located Environments. In *CSCW 2004 Workshop*. Chicago, IL.

Intille, S. S., Rondoni, J., Kukla, C., Ancona, I., & Bao, L. (2003). A context-aware experience sampling tool. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems (CHI EA '03)* (pp. 972–973). New York, NY, USA: ACM Press. doi:10.1145/765891.766101

Isbister, K., Hook, K., Laaksolahti, J., & Sharp, M. (2007). The sensual evaluation instrument: Developing a trans-cultural self-report measure of affect. *International Journal of Human-Computer Studies*, *65*(4), 315–328. doi:10.1016/j.ijhcs.2006.11.017

ISO. (1998). ISO 9241-11:1998: Ergonomic requirements for office work with visual display terminals (VDTs)-Part 11: guidance on usability.
ISO. (2006). ISO/IEC 25062 : Software Engineering —Software product Quality Requirements and Evaluation (SQuaRE —Common Industry Format (CIF) for Usability Test Reports).
ISO. (2008). ISO DIS 9241-210: Ergonomics of human system interaction. Geneva: International Standards Organization.
ISO 9241-210:2010, Ergonomie de l'interaction homme-système — Partie 210: Conception centrée sur l'opérateur humain pour les systèmes interactifs. (2010). Geneve: International Organization for Standardization.
Isomursu, M., Kuutti, K., & Väinämö, S. (2004). Experience Clip : Method for User Participation and Evaluation of Mobile Concepts. In A. Press (Ed.), *Proceedings of the 8th Conference on Participatory Design - Artful integration: interweaving Media, Materials and Practices (PDC04)* (pp. 83–92). New York. doi:10.1145/1011870.1011881
Ivory, J. D., & Magee, R. G. (2009). You can't take it with you? Effects of handheld portable media consoles on physiological and psychological responses to video game and movie content. *Cyberpsychology & Behavior : The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, *12*(3), 291–7. doi:10.1089/cpb.2008.0279
Ivory, M., & Hearst, M. (2001). The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys*, *33*(4), 470–516. doi:10.1145/503112.503114
Izard, C. E. (1972). *Patterns of emotion: A new analysis of anxiety and depression*. New York: Academic Press.
Jacko, J., & Sears, A. (2002). *The Human-Computer Interaction Handbook: Fondamentals, Evolving Technologies and Emerging Applications*. New Jersey: Lawrence Erlbaum & Associates.
Jackson, M. H. (2009). The Mash-Up: A New Archetype for Communication. *Journal of Computer-Mediated Communication*, *14*(3), 730–734. doi:10.1111/j.1083-6101.2009.01463.x
Jacob, R. J. K., & Karn, K. S. (2003). Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises. In & H. D. (eds. . . In R. Radach, J. Hyona (Ed.), *The mind's eye: cognitive and applied aspects of eye movement research* (pp. 573–605). Boston: North-Holla.
Jacovi, M., Soroka, V., Gilboa-Freedman, G., Ur, S., Shahar, E., & Marmasse, N. (2006). The chasms of CSCW : a citation graph analysis of the CSCW conference. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work - CSCW '06* (p. 289). New York, New York, USA: ACM Press. doi:10.1145/1180875.1180920
Jain, J., Evans, M., & Vinayagamoorthy, V. (2013). Exploring and enhancing the user experience for TV. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13* (p. 3187). New York, New York, USA: ACM Press. doi:10.1145/2468356.2479643
Jakobson, R. (1963). *Essais de linguistique générale* (Minuit.). Paris, France.
Jana, R. (2011). Facebook's Design Strategy: A Status Update, Behind the scenes with the team that's redefining human connection. *Design Mind*, *14*.
Jäncke, L. (1994). An EMG investigation of the coactivation of facial muscles during the presentation of affect-laden stimuli. *Journal of Psychophysiology*, *8*, 1–10.
Jean-Marc Robert, & Brangier, E. (2009). What Is Prospective Ergonomics? A Reflection and a Position on the Future of Ergonomics. In B.-T. Karsh (Ed.), *Ergonomics and Health Aspects of Work with Computers* (Vol. 5624). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-02731-4_19
Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. (1991). User interface evaluation in the real world. In *Proceedings of the SIGCHI conference on Human factors in computing systems Reaching through technology - CHI '91* (pp. 119–124). New York, New York, USA: ACM Press. doi:10.1145/108844.108862
Jenett, C. (2010). *Is game immersion just another form of selective attention? An empirical investigation of real world dissociation in computer game immersion*. University College London.
Jenkins, S., Brown, R., & Rutterford, N. (2009). Comparing Thermographic , EEG , and Subjective Measures of Affective Experience During Simulated Product Interactions. *International Journal of Design*, *3*(2), 53–65.
Jenkinson, A. (1994). Beyond segmentation. *Journal of Targeting, Measurement and Analysis of Marketing*, *1*.
Jennes, I., & Pierson, J. (2011). Audience measurement and digitalisation. In *Proceddings of the 9th international interactive conference on Interactive television - EuroITV '11* (p. 97). New York, New York, USA: ACM Press. doi:10.1145/2000119.2000138
Jennett, C., Cox, A., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., & Walton, A. (2008). Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies*, *66*(9), 641–661. doi:10.1016/j.ijhcs.2008.04.004
Jennett, C., Cox, A. L., & Cairns, P. (2009). Investigating computer game immersion and the component real world dissociation. In *Proceedings of the 27th international conference extended abstracts on Human*

factors in computing systems - CHI EA '09 (p. 3407). New York, New York, USA: ACM Press. doi:10.1145/1520340.1520494
Jensen, D. (2000). Data snooping, dredging and fishing: The dark side of data mining. *ACM SIGKDD Explorations Newsletter*, *1*(2), 52–54.
Jick, T. D. (1979). Mixing Qualitative and Quantitative Methods: Triangulation in Action. *Administrative Science Quarterly*, *24*(4), 602. doi:10.2307/2392366
Johansen, R. (1988). *Groupware: Computer support for business teams*. New York: The Free Press.
Johansen, S. A., San Agustin, J., Skovsgaard, H., Hansen, J. P., & Tall, M. (2011). Low cost vs. high-end eye tracking for usability testing. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11* (p. 1177). New York, New York, USA: ACM Press. doi:10.1145/1979742.1979744
Jordan, P. (2000). *Designing Pleasurable Products*. London, UK: Taylor & Francis.
Jordan, P. W. (1998). Human factors for pleasure in product use (1998). *Applied Ergonomics*, *29*(1), 25–33.
Jordan, P. W. (1999). Pleasure with products : Human factors for body, mind and soul. In W. S. Green & P. W. Jordan (Eds.), *Humans factors in Product Design : Current practice and future trends* (pp. 206–217). London: Taylor & Francis.
Jordan, P. W. (2002). *Designing Pleasurable Products: An Introduction to the New Human Factors*. CRC Press. Retrieved from http://www.amazon.co.uk/Designing-Pleasurable-Products-Introduction-Factors/dp/0415298873
Jung, C. G. (1907). On the Psychophysical relations of the association experiment. *Journal of Abnormal Psychology*, *1*(6), 247–255.
Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*, 441–480.
Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of Thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology*, *47*, 310–339.
Justesen, S. (2001). *—Innoversity— the dynamic relationship between innovation and diversity*.
Kahneman, D. (1973). *Attention and effort*. New York: Prentice Hall.
Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar Straus Giroux.
Kainda, R., Flechais, I., & Roscoe, A. W. (2009). Usability and security of out-of-band channels in secure device pairing protocols. In *Proceedings of the 5th Symposium on Usable Privacy and Security*.
Kallinen, K., & Ravaja, N. (2005). Effects of the rate of computer-mediated speech on emotion-related subjective and physiological responses. *Behaviour & Information Technology*, *24*(5), 365–373. doi:10.1080/01449290512331335609
Kane, M. T., & Brennan, R. L. (1977). The generalizability of class means. *Review of Educational Research*, *47*, 267–292.
Kanefsky, B., Barlow, N. G., & Gulick, V. C. (2001). Can Distributed Volunteers Accomplish Massive Data Analysis Tasks? In *Proceedings of the Lunar and Planetary Science Conference XXXII*.
Kaplan, B., & Duchon, D. (1988). Combining qualitative and quantitative methods in information systems research: a case study. *MIS Quarterly*, *12*, 571–586.
Kapoor, a, Burleson, W., & Picard, R. (2007). Automatic prediction of frustration. *International Journal of Human-Computer Studies*, *65*(8), 724–736. doi:10.1016/j.ijhcs.2007.02.003
Karapanos, E., Hassenzahl, M., & Martens, J.-B. (2008). User experience over time. In *Proceeding of the twenty-sixth annual CHI conference extended abstracts on Human factors in computing systems - CHI '08* (p. 3561). New York, New York, USA: ACM Press. doi:10.1145/1358628.1358891
Karapanos, E., Martens, J., & Hassenzahl, M. (2010). On the retrospective assessment of Users' Experiences Over Time: Memory or Actuality? In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems - CHI EA '10* (p. 4075). New York, New York, USA: ACM Press. doi:10.1145/1753846.1754105
Karapanos, E., Martens, J.-B., & Hassenzahl, M. (2009). Reconstructing Experiences through Sketching (soumis).
Karapanos, E., Zimmerman, J., Forlizzi, J., & Martens, J. (2009). User experience over time: An Initial Framework. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI '09* (p. 729). New York, New York, USA: ACM Press. doi:10.1145/1518701.1518814
Karat, C.-M., Campbell, R., & Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '92* (pp. 397–404). New York, New York, USA: ACM Press. doi:10.1145/142750.142873
Karpouzis, K. (2011). Editorial : " Signals to Signs " – Feature Extraction , Recognition , and Multimodal Fusion. In R. Cowie, C. Pelachaud, & P. Petta (Eds.), *Emotion-Oriented Systems* (pp. 65–69). Berlin, Heidelberg: Springer Berlin Heidelberg.

Karsenty, L. (2004). Enjeux, rôles et limites d'une approche ergonomique de la conception de produits. In J. Caelen (Ed.), *Le consommateur au cœur de l'innovation* (pp. 129–146). Paris: CNRS Éditions.
Kaufmann, T., Sütterlin, S., Schulz, S. M., & Vögele, C. (2011). ARTiiFACT: a tool for heart rate artifact processing and heart rate variability analysis. *Behavior Research Methods*, *43*, 1161–70. doi:10.3758/s13428-011-0107-7
Kay, L. E. (2000). *Who Wrote the Book of Life? A History of the Genetic Code*. Chicago: University of Chicago Press.
Kaye, J. J., & Sengers, P. (2007). The Evolution of Evaluation. In *Alt. CHI, proceedings of the 2007 annual conference on Human factors in computing systems - CHI'07*.
Kaye, J. "Jofish," Buie, E., Hoonhout, J., Höök, K., Roto, V., Jenson, S., & Wright, P. (2011). Panel: Designing for User Experience: Academia & Industry. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11* (p. 219). New York, New York, USA: ACM Press. doi:10.1145/1979742.1979486
Keller, J., Bless, H., Blomann, F., & Kleinböhl, D. (2011). Physiological aspects of flow experiences: Skills-demand-compatibility effects on heart rate variability and salivary cortisol. *Journal of Experimental Social Psychology*, *47*(4), 849–852. doi:10.1016/j.jesp.2011.02.004
Kelley, J. F. (1983). An empirical methodology for writing user-friendly natural language computer applications. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI '83* (pp. 193–196). New York, New York, USA: ACM Press. doi:10.1145/800045.801609
Kelly, G. (1955). *The psychology of personal constructs. Vol 1 & 2*. London, UK: Routledge.
Khawaja, M. A., Chen, F., & Marcus, N. (2014). Measuring Cognitive Load Using Linguistic Features: Implications for Usability Evaluation and Adaptive Interaction Design. *International Journal of Human-Computer Interaction*, *30*(5), 343–368. doi:10.1080/10447318.2013.860579
Kieras, D. E., & Santoro, T. P. (2004). Computational GOMS modeling of a complex team task: lessons learned. In *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04* (pp. 97–104). New York, New York, USA: ACM Press. doi:10.1145/985692.985705
Kim, J. H., Gunn, D. V, Schuh, E., Phillips, B., Pagulayan, R. J., & Wixon, D. (2008). Tracking real-time user experience (TRUE): A comprehensive instrumentation solution for complex systems. In *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08* (p. 443). New York, New York, USA: ACM Press. doi:10.1145/1357054.1357126
King, G. (2011). Ensuring the data-rich future of the social sciences. *Science*, *331*(6018), 719–721. doi:10.1126/science.1197872
Kirakowski, J. (1996). The Software Usability Measurement Inventory: Background and usage. In P. Jordan, B. Thomas, & B. Weerdmeester (Eds.), *Usability Evaluation in Industry* (pp. 169–178). London, UK: Taylor & Francis.
Kirakowski, J., & Cierlik, B. (1998). Measuring the Usability of Web Sites. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *42*(4), 424–428. doi:10.1177/154193129804200405
Kirakowski, J., & Corbett, M. (1993). SUMI: the Software Usability Measurement Inventory. *British Journal of Educational Technology*, *24*(3), 210–212. doi:10.1111/j.1467-8535.1993.tb00076.x
Kirakowski, J., & Dillon, A. (1998). *The Computer User Satisfaction Inventory (CUSI): Manual and Scoring Key*. Cork, Ireland.
Kissel, G. (1995). The effect of computer experience on subject and objective software usability measures. In *Proceedings of CHI'95* (pp. 284–285).
Kittur, A., & Kraut, R. E. (2008). Harnessing the wisdom of crowds in wikipedia. In *Proceedings of the ACM 2008 conference on Computer supported cooperative work - CSCW '08* (p. 37). New York, New York, USA: ACM Press. doi:10.1145/1460563.1460572
Kittur, A., Suh, B., Chi, E., & Pendleton, B. A. (2007). He says, she says: Conflict and coordination in Wikipedia. In *CHI 2007* (pp. 453–462). ACM Press.
Kivikangas, J. M. (2006). *Psychophysiology of flow experience: An explorative study.* Helsinki.
Kivikangas, J. M., Ekman, I., Chanel, G., Järvelä, S., Salminen, M., Cowley, B., … Ravaja, N. (2010). Review on psychophysiological methods in game research. In *Proc. of 1st Nordic DiGRA* (Vol. 2010). Retrieved from http://www.digra.org/dl/db/10343.06308.pdf
Klein, E., Tellefsen, T., & Herskovitz, P. (2007). The use of group support systems in focus groups: Information technology meets qualitative research. *Computers in Human Behavior*, *23*(5), 2113–2132. doi:10.1016/j.chb.2006.02.007
Klemperer, P. (2006). *Network Effects and Switching Costs: Two Short Essays for the New Palgrave*.
Kline, S. J., & Rosenberg, N. (1986). An overview of Innovation. In R. Landau & N. Rosenberg (Eds.), *The Positive Sum Strategy : Harnessing Technology for Economic Growth* (pp. 275–305). Washington D.C.: National Academy Press.

Knight, K. (2011, January). Responsive Web Design: What It Is and How To Use It. *Smashing Magazine*. Retrieved from http://www.smashingmagazine.com/2011/01/12/guidelines-for-responsive-web-design/

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, *22*(4-5), 441–504. doi:10.1007/s11257-011-9118-4

Knoll, A., Wang, Y., Chen, F., & Xu, J. (2011). Measuring Cognitive Workload with Low-Cost Electroencephalograph. In P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, & M. Winckler (Eds.), *Human-Computer Interaction – INTERACT 2011* (Vol. 6949, pp. 568–571). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-23768-3

Kohlbecher, S., Bardinst, S., Bartl, K., Schneider, E., Poitschke, T., & Ablassmeier, M. (2008). Calibration-free eye tracking by reconstruction of the pupil ellipse in 3D space. In *Proceedings of the 2008 symposium on Eye tracking research & applications - ETRA '08* (Vol. 1, p. 135). New York, New York, USA: ACM Press. doi:10.1145/1344471.1344506

Konstan, J. a., & Riedl, J. (2012). Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, *22*(1-2), 101–123. doi:10.1007/s11257-011-9112-x

Korhonen, H., Paavilainen, J., & Saarenpää, H. (2009). Expert review method in game evaluations: comparison of two playability heuristic sets. In *Proc. of MindTrek 2009* (pp. 74–81).

Kort, J., Vermeeren, A. P. O. S., & Fokker, J. E. (2007). Conceptualization and measuring UX. In E. L.-C. Law, A. Vermeeren, M. Hassenzahl, & M. Blythe (Eds.), *Towards a UX manifesto. COST294-MAUSE affiliated workshop* (pp. 57–64). Lancaster: COST.

Kortum, P., & Peres, S. C. (2014). The relationship between system effectiveness and subjective usability scores using the System Usability Scale. *International Journal of Human–Computer Interaction*, *30*, 575–584.

Kortum, P. T., & Bangor, A. (2013). Usability Ratings for Everyday Products Measured With the System Usability Scale. *International Journal of Human-Computer Interaction*, *29*(2), 67–76. doi:10.1080/10447318.2012.681221

Koza, M. P., & Lewin, A. Y. (1998). The co-evolution of Strategic Alliances. *Organization Science*, *9*(3), 255–264.

Kramer, A. D. I. (2010). An unobtrusive behavioral model of "gross national happiness." In *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10* (pp. 287–290). doi:10.1145/1753326.1753369

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(24), 8788–8790. doi:10.1073/pnas.1320040111

Kraut, R., Kiesler, S., Boneva, B., Cummings, J., Helgeson, V., & Crawford, A. (2002). Internet Paradox Revisited. *Journal of Social Issues*, *58*(1), 49–74. doi:10.1111/1540-4560.00248

Kraut, R., Patterson, M., Lundmark, V., Kiesler, S., Mukophadhyay, T., & Scherlis, W. (1998). Internet paradox: A social technology that reduces social involvement and psychological well-being? *American Psychologist*, *53*(9), 1017–1031. doi:10.1037/0003-066X.53.9.1017

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Kujala, S., & Nurkka, P. (2009). Product symbolism in designing for user experience. In *Proceedings of the 4th International Conference on Designing Pleasurable Products and Interface (DPPI 09)*. Compiegne.

Kujala, S., & Nurkka, P. (2012). Sentence Completion for Understanding Users and Evaluating User Experience. *International Journal of Design*, *6*(3), 15–25.

Kujala, S., Roto, V., Väänänen-Vainio-Mattila, K., Karapanos, E., & Sinnelä, A. (2011). UX Curve: A method for evaluating long-term user experience. *Interacting with Computers*, *23*(5), 473–483. doi:10.1016/j.intcom.2011.06.005

Kujala, S., Walsh, T., Nurkka, T., & Crisan, M. (2013). Sentence Completion for Understanding Users and Evaluating User Experience. *Interacting with Computers*, *26*(3), 238–255.

Kukolja, D., Popović, S., Horvat, M., Kovač, B., & Ćosić, K. (2014). Comparative analysis of emotion estimation methods based on physiological measurements for real-time applications. *International Journal of Human-Computer Studies*, *72*(10-11), 717–727. doi:10.1016/j.ijhcs.2014.05.006

Kuniavsky, M. (2003). *Observing the user experience a practitioner's guide to user research*. San Francisco: Morgan Kaufmann Publishers.

Kupermintz, H. (2003). Lee J. Cronbach's contributions to educational psychology. In B. J. Zimmerman & D. H. Schunk (Eds.), *Educational Psychology: A Century of Contributions* (pp. 289–302).

Kurosu, M., & Kashimura, K. (1995). Apparent usability vs. inherent usability. In *Conference companion on Human factors in computing systems - CHI '95* (pp. 292–293). New York, New York, USA: ACM Press. doi:10.1145/223355.223680

Lallemand, C., Gronier, G., & Koenig, V. (2015). User experience : A concept without consensus? Exploring practitioners' perspectives through an international survey. *Computers in Human Behavior*, *43*, 35–48. doi:10.1016/j.chb.2014.10.048

Landay, J. A., & Myers, B. A. (1996). Sketching storyboards to illustrate interface behaviors. In *Conference companion on Human factors in computing systems common ground - CHI '96* (pp. 193–194). New York, New York, USA: ACM Press. doi:10.1145/257089.257257

Lang, P. J. (1980). Behavioral treatment and bio-behavioral assessment : computer applications. In J. B. Sidowski, J. H. Johnson, & T. A. Williams (Eds.), *Technology in mental health care delivery systems* (pp. 119–137). Norwood, NJ: Ablex.

Lang, P. J. (1994). The varieties of emotional experience: a meditation on James–Lange theory. *Psychological Review*, *101*(2), 211–221.

Lang, P. J., & Bradley, M. M. (2010). Emotion and the motivational brain. *Biological Psychology*, *84*(3), 437–450. doi:10.1016/j.biopsycho.2009.10.007

Lang, P. J., Greenwald, M. K., Bradley, M. M., & Hamm, A. O. (1993). Looking at pictures: affective, facial, visceral, and behavioural reactions. *Psychophysiology*, *30*, 261–273.

Larsen, J. T., Norris, C. J., & Cacioppo, J. T. (2003). Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii. *Psychophysiology*, *40*(5), 776–785.

Larsen, K. R., Allen, G., Vance, A., & Eargle, D. (2014). Theories Used in IS Research Wiki. Retrieved August 22, 2014, from http://istheory.byu.edu

Latour, B. (2009). Tarde's idea of quantification. In M. Candea (Ed.), *The Social After Gabriel Tarde: Debates and Assessments* (pp. 145–162). London: Routledge.

Latulipe, C., Carroll, E. A., & Lottridge, D. (2011). Love, hate, arousal and engagement: Exploring Audience Responses to Performing Arts. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11* (p. 1845). New York, New York, USA: ACM Press. doi:10.1145/1978942.1979210

Laurans, G., & Desmet, P. M. A. (2006). Using self-confrontation to study user experience : A new approach to the dynamic measurement of emotions while interacting with products. In P. M. A. Desmet, M. A. Karlsson, & J. van Erp (Eds.), *Proceedings of the 5th international conference on Desing and Emotion (Design & Emotion 2006)*. Gothenburg, Sweden: Chalmers University.

Laviea, T., & Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. *International Journal of Human-Computer Studies*, *60*(3), 269–298. doi:10.1016/j.ijhcs.2003.09.002

Law, E. L.-C. (2011). The measurability and predictability of user experience. *Proceedings of the 3rd ACM SIGCHI Symposium on Engineering Interactive Computing Systems - EICS '11*, 1. doi:10.1145/1996461.1996485

Law, E. L.-C., & Abrahão, S. (2014). Interplay between User Experience (UX) evaluation and system development. *International Journal of Human-Computer Studies*, *72*(6), 523–525. doi:10.1016/j.ijhcs.2014.03.003

Law, E. L.-C., Bevan, N., Cockton, G., Christou, G., Springett, M., Lárusdóttir, M., … Jokela, T. (2008). Proceedings of the International Workshop on meaningful measures : Valid Useful User Experience Measurement (VUUM). In *Proceedings of the International Workshop on meaningful measures : Valid Useful User Experience Measurement (VUUM)*. Reykjavik, Iceland.

Law, E. L.-C., Hvannberg, E. T., & Hassenzahl, M. (2006). Workshop on User Experience - Towards a Unified View. In *conjunction with NordiCHI'06*. Oslo, Norway. Retrieved from http://www.cost294.org/

Law, E. L.-C., Roto, V., Hassenzahl, M., Vermeeren, A. P. O. S., & Kort, J. (2009). Understanding, scoping and defining user experience: A Survey Approach. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI '09* (p. 719). New York, New York, USA: ACM Press. doi:10.1145/1518701.1518813

Law, E. L.-C., Roto, V., Vermeeren, A. P. O. S., Kort, J., & Hassenzahl, M. (2008). Towards a shared definition of user experience. In *Proceeding of the twenty-sixth annual CHI conference extended abstracts on Human factors in computing systems - CHI '08* (p. 2395). New York, New York, USA: ACM Press. doi:10.1145/1358628.1358693

Law, E. L.-C., & van Schaik, P. (2010). Modelling user experience – An agenda for research and practice. *Interacting with Computers*, *22*(5), 313–322. doi:10.1016/j.intcom.2010.04.006

Law, E. L.-C., van Schaik, P., & Roto, V. (2014). Attitudes towards user experience (UX) measurement. *International Journal of Human-Computer Studies*, *72*(6), 526–541. doi:10.1016/j.ijhcs.2013.09.006

Law, E. L.-C., Vermeeren, A., Hassenzahl, M., & Blythe, M. (2007). *Towards a UX Manifesto: COST294-MAUSE affiliated workshop*. (E. Law, A. Vermeeren, M. Hassenzahl, & M. Blythe, Eds.) (p. 76 pages.). Lancaster, UK.

Le Moigne, J. L. (1995). *Les Epistémologies constructivistes*. Paris: Presses universitaires de France.

Lee, C. J., Jang, C.-Y. I., Chen, T. D., Wetzel, J., Shen, Y.-T. B., & Selker, T. (2006). Attention Meter: A Vision-based Input Toolkit for Interaction Designers. In *CHI '06 extended abstracts on Human factors in computing systems - CHI EA '06* (p. 1007). New York, New York, USA: ACM Press. doi:10.1145/1125451.1125644

Lee, S. M., Olson, D. L., & Trimi, S. (2012). Co-innovation: convergenomics, collaboration, and co-creation for organizational values. *Management Decision*, *50*(5), 817–831.

Lee, S., Yoon, B., & Park, Y. (2009). An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation*, *29*(6-7), 481–497. doi:10.1016/j.technovation.2008.10.006

Lee, W., Chiu, Y. T., Liu, C., & Chen, C. (2011). Assessing the effects of consumer involvement and service quality in a self-service setting. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *21*(5), 504–515. doi:10.1002/hfm.20253

Léger, L., Fouquereau, N., Levillain, F., & Tijus, C. (2009). *Projet SP3 : Méthodes en ergonomie des interfaces* (pp. 44–97). Paris.

Léger, P. M., Davis, F. D., Cronan, T. P., & Perret, J. (2014). Neurophysiological correlates of cognitive absorption in an enactive training context. *Computers in Human Behavior*, *34*, 273–283. doi:10.1016/j.chb.2014.02.011

Leonard-Barton, D. (1992). Core Capabilities and Core Rigidities: A Paradox in Managing New Product Development. *Strategic Management Journal*, *13*, 111–125.

Leroi-Gourhan, A. (1964). *Le geste et la parole*. Paris: Albin Michel.

Lesca, H. (1996). Veille stratégique : Comment sélectionner les informations pertinentes ? Concepts, méthodologie, expérimentation, résultats. In *5ème conférence internationale de management stratégique*.

Leskovec, J., & Faloutsos, C. (2006). Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 631–636). doi:10.1145/1150402.1150479

Leslie, G., Ojeda, A., & Makeig, S. (2014). Measuring musical engagement using expressive movement and EEG brain dynamics. *Psychomusicology: Music, Mind, and Brain*, *24*(1), 75–91.

Levine, R. (2000). *The cluetrain manifesto: The end of business as usual*. Cambridge, Mass.: Perseus Books.

Lewis, C. H. (1982). *Using the "Thinking Aloud" Method In Cognitive Interface Design*.

Lewis, J. (2002). Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies. *International Journal of Human-Computer Interaction*, *14*(3), 463–488. doi:10.1207/S15327590IJHC143&4_11

Lewis, J. R. (1991). Psychometric evaluation of an after-scenario questionnaire for computer usability studies. *ACM SIGCHI Bulletin*, *23*(1), 78–81. doi:10.1145/122672.122692

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, *7*(1), 57–78. doi:10.1080/10447319509526110

Lewis, J. R. (2006). Sample sizes for usability tests: mostly math, not magic. *Interactions*, *13*(8), 29–33.

Lewis, J. R. (2012). Usability testing. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (4th ed., pp. 1267–1312). New York, NY: Wiley.

Lewis, J. R. (2014). Usability: Lessons Learned … and Yet to Be Learned. *International Journal of Human-Computer Interaction*, *30*(9), 663–684. doi:10.1080/10447318.2014.930311

Lewis, J. R., Henry, S. C., & Mack, R. L. (1990). Integrated office software benchmarks : A case study. In *Proceedings of the 3rd IFIP Conference on Human–Computer Interaction, INTERACT '90* (pp. 337–343). Cambridge, UK: Elsevier Science.

Lewis, J. R., & Sauro, J. (2006). When 100% Really Isn't 100%: Improving the Accuracy of Small-Sample Estimates of Completion Rates. *Journal of Usability Studies*, *1*(3), 136–150.

Lewis, J. R., & Sauro, J. (2009). The factor structure of the System Usability Scale. In M. Kurosu (Ed.), *Human centered design* (pp. 94–103). Heidelberg, Germany: Springer-Verlag.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE : when there's no time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13* (pp. 2099–2102). New York, NY, USA,: ACM. doi:10.1145/2470654.2481287

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Measuring Perceived Usability: The SUS, UMUX-LITE, and AltUsability. *International Journal of Human-Computer Interaction*, *31*(8), 496–505. doi:10.1080/10447318.2015.1064654

Licklider, J. C. R., & Taylor, R. W. (1968). The Computer as a Communication Device. *Science and Technology*, *76*(2), 1–3.

Lin, H. X., Choong, Y.-Y., & Salvendy, G. (1997). A proposed index of usability: A method for comparing the relative usability of different software systems. *Behaviour & Information Technology*, *16*(4-5), 267–277. doi:10.1080/014492997119833

Lin, X. (2006). Active layout engine: Algorithms and applications in variable data printing. *Computer-Aided Design*, *38*(5), 444–456.
Linacre, J. M. (1989). *Many-facet Rasch Measurement*. Chicago: MESA Press.
Lindgaard, G., & Kirakowski, J. (2013). Introduction to the special issue: The tricky landscape of developing rating scales in HCI. *Interacting with Computers*, *25*(4), 271–277.
Linstone, H. A., & Turoff, M. (1975). *The Delphi method : Techniques and applications*. Reading, MA: Addison-Wesley Publishing.
Lipovetsky, G., & Serroy, J. (2013). *L'esthétisation du monde: Vivre à l'âge du capitalisme artiste*. Paris, France: Gallimard.
Liu, J., Li, J., Feng, L., Li, L., Tian, J., & Lee, K. (2014). Seeing Jesus in toast: Neural and behavioral correlates of face pareidolia. *Cortex*, *53*(0), 60–77. doi:http://dx.doi.org/10.1016/j.cortex.2014.01.013
Loewenstein, G., & Lerner, J. S. (2003). The role of affect in decisionmaking. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.) (Eds.), *Handbook of affective science* (pp. 619 – 642). New York: Oxford University Press.
Loftus, E. F., Schooler, J. W., Boone, S. M., & Kline, D. (1987). Time went by so slowly : Overestimation of event duration by males and females. *Applied Cognitive Psychology*, *1*, 3–13. doi:10.1002/acp.1778
Lombard, M., & Ditton, T. (1997). At the Heart of It All : The Concept of Presence. *Journal of Computer-Mediated Communication*, *3*(2).
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Menlo Park, CA: Addison Wesley.
Lorenz, A., & Oppermann, R. (2009). Mobile health monitoring for the elderly: Designing for diversity. *Pervasive and Mobile Computing*, *5*(5), 478–495. doi:10.1016/j.pmcj.2008.09.010
Lottridge, D. (2009). Evaluating Human Computer Interaction through Self-rated Emotion. In T. Gross, J. Gulliksen, P. Kotzé, L. Oestreicher, P. Palanque, R. O. Prates, & M. Winckler (Eds.), *Human-Computer Interaction – INTERACT 2009* (Vol. 5727, pp. 860–863). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-03658-3
Lowgren, J. (2013). Interaction Design - brief intro. In M. Soegaard & R. F. Dam (Eds.), *The Encyclopedia of Human-Computer Interaction, 2nd Ed.* Aarhus, Danmark: The Interaction Design Foundation.
Lubart, T. (2003). *Psychologie de la créativité*. Paris: Armand Colin.
Lubinski, D. (2006). Ability tests. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 101–114). Washington, DC: American Psychological Association.
Lucas, P. (2000). Pervasive information access and the rise of human-information interaction. In *CHI '00 extended abstracts on Human factors in computing systems - CHI '00* (p. 202). New York, New York, USA: ACM Press. doi:10.1145/633292.633405
Lucas, R. E., & Baird, B. M. (2006). Global self-assessment. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 29–42). Washington, DC: American Psychological Association.
Lucero, A. (2009). *Co-designing interactive spaces for and with designers: supporting mood-board making*. Eindhoven University of Technology (TU/e), Netherlands.
Lund, A. M. (2001). Measuring usability with the USE questionnaire. Usability Interface. *Usability & User Experience*, *8*(2).
Luria, A. R. (1976). *Cognitive Development, Its Cultural and Social Foundations*. Cambridge, Mass.: Harvard University Press.
MacDonald, C. M., & Atwood, M. E. (2013). Changing perspectives on evaluation in HCI. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13* (pp. 1969–1978). New York, New York, USA: ACM Press. doi:10.1145/2468356.2468714
MacKenzie, S. B. (2003). The Dangers of Poor Construct Conceptualization. *Journal of the Academy of Marketing Science*, *31*(3), 323–326.
MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct Measurement and Validation Procedures in MIS and Behavioral Research: Integrating New and Existing Techniques. *MIS Quarterly*, *35*(2), 293–334. Retrieved from http://aisel.aisnet.org/misq/vol35/iss2/5/
Madsen, M., & Gregor, S. (2000). Measuring Human-Computer Trust. In G. Gable & M. Viatle (Eds.), *Proceedings of 11th Australasian Conference on Information Systems* (pp. 6–8).
Magnusson, P. R. (2009). Exploring the Contributions of Involving Ordinary Users in Ideation of Technology-Based Services. *Journal of Product Innovation Management*, *26*(5), 578–593. doi:10.1111/j.1540-5885.2009.00684.x
Mahlke, S. (2008). *User Experience of Interaction with Technical Systems: Theories, Methods, Empirical Results, and Their Application to the Development of Interactive Systems*. Technischen Universität Berlin.
Mäkelä, A., & Fulton Suri, J. (2001). Supporting users' creativity: Design to induce pleasurable experiences. In *Proceedings of the International Conference on Affective Human Factors Design* (pp. 387–394).

Malone, T. W. (1982). Heuristics for designing enjoyable user interfaces. In *Proceedings of the 1982 conference on Human factors in computing systems - CHI '82* (pp. 63–68). New York, New York, USA: ACM Press. doi:10.1145/800049.801756

Mandryk, R., & Atkins, M. (2007). A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International Journal of Human-Computer Studies*, *65*(4), 329–347. doi:10.1016/j.ijhcs.2006.11.011

Mandryk, R. L., Inkpen, K. M., & Calvert, T. W. (2006). Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & Information Technology*, *25*(2), 141–158. doi:10.1080/01449290500331156

Maniak, R., & Midler, C. (2008). Shifting from co-development to co-innovation. *International Journal of Automotive Technology & Management*, *8*(4), 449–468.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*.

Marache-francisco, C., & Brangier, E. (2013). Process of Gamification From The Consideration of Gamification To Its Practical Implementation. In L. Berntzen & S. Böhm (Eds.), *CENTRIC 2013: The Sixth International Conference on Advances in Human oriented and Personalized Mechanisms, Technologies, and Services* (pp. 126–131). Red Hook: Curran Associates, Inc. doi:10.1007/978-3-642-39241-2_61

Marchionini, G. (2008). Human–information interaction research and development. *Library & Information Science Research*, *30*(3), 165–174. doi:10.1016/j.lisr.2008.07.001

Marcotte, E. (2011). *Responsive Web design*. Paris: Eyrolles.

Marie, N. (2014). *Linked data based exploratory search*. Université de Nice Sophia-Antipolis.

Marie, N., Gandon, F., Ribière, M., & Rodio, F. (2013). Discovery hub : on-the-fly linked data exploratory search. In *Proceedings of the 9th International Conference on Semantic Systems - I-SEMANTICS '13* (pp. 17–24). doi:10.1145/2506182.2506185

Marsh, H. W. (1994). Sport motivation orientations: Beware of jingle–jangle fallacies. *Journal of Sport & Exercise Psychology*, *16*, 365–380.

Marshalek, B., Lahman, D. F., & Snow, R. E. (1983). The complexity continuum in radix and hierarchical models of intelligence. *Intelligence*, *4*, 107–127.

Martin, J. (1973). *Design of man-computer dialogues.* Englewood Cliffs, N.J: Prentice-Hall.

Maslow, A. H. (1954). *Motivation and personality*. New York: Harper.

Mason, J. (1996). *Qualitative researching*. London: Sage.

Matias Kivikangas, J., Nacke, L., & Ravaja, N. (2011). Developing a triangulation system for digital game events, observational video, and psychophysiological data to study emotional responses to a virtual character. *Entertainment Computing*, *2*(1), 11–16. doi:10.1016/j.entcom.2011.03.006

Mauboussin, M. J. (2007). The Wisdom and Whims of the Collective. *CFA Institute Conference Proceedings Quarterly*, *24*(4), 1–8. doi:10.2469/cp.v24.n4.4934

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin Harcourt.

Mayhew, D. J. (2009). Requirements specifications within the usability engineering lifecycle. In A. Sears & J. A. Jacko (Eds.), *Human Computer Interaction : Development Process* (pp. 23–32). New York: CRC Press.

McCarthy, J., & Wright, P. (2004). Technology as experience. *Interactions*, *11*(5), 42. doi:10.1145/1015530.1015549

McClure, D. (2007). Product Marketing for Pirates: AARRR! (aka Startup Metrics for Internet Marketing & Product Management). *500hats.typepad.com*. Retrieved from http://500hats.typepad.com/500blogs/2007/06/internet-market.html

McCulloch, W. S., & Pitts, W. H. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, *5*, 115–133.

McGee, M. (2003). Usability magnitude estimation. In *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting* (pp. 691–695). Santa Monica, CA: HFES.

McGrath, J. E. (1981). Dilemmatics : The study of research choices and dilemmas. *American Behavioral Scientist*, *25*(2), 154–179.

Mcintire, L., Goodyear, C., Bridges, N., Mckinley, R. A., Merritt, M., Griffin, K., & Mcintire, J. (2011). *Eye-Tracking: An Alternative Vigilance Detector*.

McNee, S. M., Riedl, J., & Konstan, J. A. (2006). Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *proceeding of CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06)* (pp. 1097–1101). New York, NY, USA: ACM.

Medlock, C. M., Wixon, D., Terrano, M., Romero, R. L., & Fulton, B. (2002). Using the RITE method to improve products: a definition and a case study. In *Proceedings of UPA'02*.

Mehl, M. R. (2006). Quantitative text analysis. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 141–156). Washington, DC: American Psychological Association.

Mehrabian, A., & de Wetter, R. (1987). Experimental test of an emotion-based approach to fitting brand names to products. *Journal of Applied Psychology*, *72*(1), 125–130. doi:10.1037/0021-9010.72.1.125
Meijman, T., Zijlstra, F., Kompier, M., & Mulders, H. (1986). *The measurement of perceived effort (Report)*. University of Groningen, Department of Occupational Psychology, Biological Center, Haren, The Netherland.
Meister, D. (1985). *Behavioral Analysis and Measurement Methods*. New York: Wiley.
Menzel, H. C., Aaltio, I., & Ulijn, J. M. (2007). On the way to creativity : Engineers as intrapreneurs in organizations. *Technovation*, *27*(12), 732–743.
Michaud, Y. (2003). *L'art à l'état gazeux. Essai sur le triomphe de l'esthétique*. Paris, France: Hachette Littératures.
Michaud, Y. (2013). *Le nouveau luxe: expériences, arrogance, authenticité*. Paris: Stock.
Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., & Gray, M. K. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*, 176–182.
Middleton, H. C., Sharma, A., Agouzoul, D., Sahakian, B. J., & Robbins, T. W. (1999). Contrasts between the cardiovascular concomitants of tests of planning and attention. *Psychophysiology*, *36*(5), 610–618.
Midler, C. (2003). *L'auto qui n'existait pas : Management des projets et transformation de l'entreprise*. Paris: InterEditions.
Midler, C., Maniak, R., & Beaume, R. (2007). Du co-développement à la co-innovation-Analyse empirique des coopérations verticales en conception innovante. In *15th GERPISA International Colloquium*. Paris.
Miller, A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review*, *63*, 81–97.
Mingers, J. (2001). Combining research methods: towards a pluralistic methodology. *Information Systems Research*, *12*, 240–259.
Mintzberg, H. (2004). *Le management : voyage au centre des organisations*. Paris: Éditions d'Organisation.
Mirizzi, R., Di Noia, T., Ragone, A., Ostuni, V. C., & Di Sciascio, E. (2012). Movie recommendation with DBpedia. In G. Amati, C. Carpineto, & G. Semeraro (Eds.), *Proceedings of the Third Italian Information Retrieval Workshop (IIR 2012)* (pp. 101–112). Bari Aldo Moro, Italy.
Mirza-Babaei, P., & McAllister, G. (2010). Using physiological measures in conjunction with other usability approaches for better understanding of the player's gameplay experiences. In *Game Research Methods* (pp. 1–15).
Mirza-Babaei, P., Nacke, L., Fitzpatrick, G., White, G., McAllister, G., & Collins, N. (2012). Biometric storyboards : Visualising Game User Research Data. In *Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts - CHI EA '12* (p. 2315). New York, New York, USA: ACM Press. doi:10.1145/2212776.2223795
Mishra, B., Fernandes, S. L., Abhishek, K., Alva, A., Shetty, C., & Ajila, C. (2015). Facial expression recognition using feature based techniques and model based techniques: A survey. In *2nd International Conference on Electronics and Communication Systems (ICECS)* (pp. 589–594). IEEE.
Miyake, S., Yamada, S., Shoji, T., Takae, Y., Kuge, N., & Yamamura, T. (2009). Physiological responses to workload change. A test/retest examination. *Applied Ergonomics*, *40*(6), 987–996. doi:10.1016/j.apergo.2009.02.005
Model, D., & Eizenman, M. (2012). A general framework for extension of a tracking range of user-calibration-free remote eye-gaze tracking systems. In *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12* (Vol. 1, p. 253). New York, New York, USA: ACM Press. doi:10.1145/2168556.2168609
Molich, R., & Dumas, J. S. (2008). Comparative usability evaluation (CUE-4). *Behaviour & Information Technology*, *27*(3), 263–281. doi:10.1080/01449290600959062
Molin, L. (2004). Wizard-of-Oz prototyping for co-operative interaction design of graphical user interfaces. In *Proceedings of the third Nordic conference on Human-computer interaction - NordiCHI '04* (pp. 425–428). New York, New York, USA: ACM Press. doi:10.1145/1028014.1028086
Moneta, G. B. (2012). On the measurement and conceptualization of flow. In S. Engeser (Ed.), *Advances in Flow Research* (pp. 23–50). Boston, MA: Springer US. doi:10.1007/978-1-4614-2359-1
Money, A. G., & Agius, H. (2009). Analysing user physiological responses for affective video summarisation. *Displays*, *30*(2), 59–70. doi:10.1016/j.displa.2008.12.003
Morin, E. (1973). *Le paradigme perdu : la nature humaine*. Paris: Seuil.
Morin, E. (1986). *La Connaissance de la Connaissance (La Méthode: tome 3)*. Paris: Seuil.
Morin, E. (1995). La stratégie de reliance pour l'intelligence de la complexité. *Revue Internationale de Systémique*, *9*(2).
Morin, E. (2003). *L'humanité de l'humanité (La méthode : Tome 5)*. Paris: Seuil.
Mota, S., & Picard, R. W. (2003). Automated posture analysis for detecting learner's interest level. In *CVPRW'03: Computer Vision and Pattern Recognition Workshop*.

Moulier-Boutang, Y. (2007). *Le capitalisme cognitif : La nouvelle grande transformation*. Paris: Éd. Amsterdam.
Moulier-Boutang, Y. (2010). *L'abeille et l'économiste*. Paris: Carnets Nord.
Muckler, F. A., & Seven, S. A. (1992). Selecting performance measures: "'objective'" versus "'subjective'" measurement. *Human Factors*, *34*(4), 441–455.
Mukherjee, A., Liu, B., & Glance, N. (2012). Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on World Wide Web - WWW '12* (p. 191). New York, New York, USA: ACM Press. doi:10.1145/2187836.2187863
Muller, M. J., & Kuhn, S. (1993). Participatory design. *Communications of the ACM*, *36*(6), 24–28. doi:10.1145/153571.255960
Mumford, E. (1971). A comprehensive method for handling the human problems of computer introduction. In *IFIP Congress, 2* (pp. 918–923).
Mutlu, B., Krause, A., Forlizzi, J., Guestrin, C., & Hodgins, J. (2007). Robust, low-cost, non-intrusive sensing and recognition of seated postures. In *Proceedings of the 20th annual ACM symposium on User interface software and technology - UIST '07* (Vol. 4, p. 149). New York, New York, USA: ACM Press. doi:10.1145/1294211.1294237
Naber, M., Frässle, S., Rutishauser, U., & Einhäuser, W. (2013). Pupil size signals novelty and predicts later retrieval success for declarative memories of natural scenes. *Journal of Vision*, *13*(2). doi:10.1167/13.2.11
Nacke, L. E. (2009). *Affective Ludology : Scientific Measurement of User Experience in Interactive Entertainment*. Blekinge Institute of Technology.
Nacke, L. E., Grimshaw, M. N., & Lindley, C. a. (2010). More than a feeling: Measurement of sonic user experience and psychophysiology in a first-person shooter game. *Interacting with Computers*, *22*(5), 336–343. doi:10.1016/j.intcom.2010.04.005
Nacke, L., & Lindley, C. (2008). Boredom, Immersion, Flow - a Pilot Study Investigating Player Experience. In *Proceedings of the IADIS Gaming 2008: Design for Engaging Experience and Social Interaction* (pp. 103–107). IADIS Press.
Nagamachi, M. (1995). Kansei engineering : a new ergonomic consumer-oriented technology for product development. *International Journal of Industrial Ergonomics*, *15*, 3–11.
Nagamachi, M. (2002). Kansei Engineering as a Powerful Consumer-Oriented Technology for Product Development. *Applied Ergonomics*, *33*(3), 289–294.
Nahin, a. F. M. N. H., Alam, J. M., Mahmud, H., & Hasan, K. (2014). Identifying emotion by keystroke dynamics and text pattern analysis. *Behaviour & Information Technology*, *33*(9), 987–996. doi:10.1080/0144929X.2014.907343
Nakano, Y. I., & Ishii, R. (2010). Estimating User 's Engagement from Eye-gaze Behaviors in Human-Agent Conversations. In *Proceedings of the 15th international conference on Intelligent user interfaces - IUI '10* (pp. 139–148). New York, New York, USA: ACM Press. doi:10.1145/1719970.1719990
Napoli, P. (2010). *Audience evolution. New technologies and the transformation of media audiences*. New York: Columbia University Press.
Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., & Giovannini, E. (2005). *Handbook on constructing composite indicators*. OECD.
Neale, D. C., & Carroll, J. M. (1999). Multi-faceted evaluation for complex, distributed activities. In C. M. Hoadley & J. Roschelle (Eds.), *Proceedings of the 1999 Conference on Computer Support for Collaborative Learning (CSCL '99)*. Stanford, CA, USA: International Society of the Learning Sciences.
Nelson, T. H. (1965). Complex information processing : a file structure for the complex, the changing and the indeterminate. In *Proceedings of the 1965 20th national conference on* - (pp. 84–100). New York, New York, USA: ACM Press. doi:10.1145/800197.806036
Nelson, T. H. (1973). A conceptual framework for man-machine everything. In *Proceedings of the June 4-8, 1973, national computer conference and exposition on - AFIPS '73* (pp. 21–26). New York, New York, USA: ACM Press. doi:10.1145/1499586.1499776
Nemery, A., Brangier, E., & Kopp, S. (2010). Proposition d'une grille de critères d'analyse ergonomiques des formes de persuasion interactive. In *Conference Internationale Francophone sur I'Interaction Homme-Machine - IHM '10* (pp. 153–156). New York, New York, USA: ACM Press. doi:10.1145/1941007.1941034
Némery, A., Brangier, E., & Kopp, S. (2011). First Validation of Persuasive Criteria for Designing and Evaluating the Social Influence of User Interfaces: Justification of a Guideline. In A. Marcus (Ed.), *Design, User Experience, and Usability. Theory, Methods, Tools and Practice* (Vol. 6770, pp. 616–624). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-21708-1
Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures : Issues and applications*. Thousand Oaks, CA: Sage Publications, Inc.

Neuman, W. R. (1990). *Beyond HDTV: Exploring subjective responses to very high definition television*. MIT, Cambridge, MA.
Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2010). EmoHeart: Conveying Emotions in Second Life Based on Affect Sensing from Text. *Advances in Human-Computer Interaction*, *2010*(2), 1–13. doi:10.1155/2010/209801
Newell, A., & Card, S. K. (1985). The prospects for psychological science in human-computer interaction. *Human-Computer Interaction*, *1*(3), 209–242.
Nielsen, J. (1989). Usability engineering at a discount. In G. Salvendy & M. Smith (Eds.), *Designing and Using Human-Computer Interfaces and Knowledge Based Systems* (pp. 394–401). Amsterdam: Elsevier Science Publishers.
Nielsen, J. (1992). Finding usability problems through heuristic evaluation. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '92* (pp. 373–380). New York, New York, USA: ACM Press. doi:10.1145/142750.142834
Nielsen, J. (1993). *Usability Engineering*. Boston, MA: Academic Press.
Nielsen, J. (1994a). Guerrilla HCI: using discount usability engineering to penetrate the intimidation barrier. In R. G. Bias & D. J. Mayhew (Eds.), *Cost-justifying usability* (pp. 245–272). Orlando, FL, USA: Academic Press, Inc..
Nielsen, J. (1994b). Usability inspection methods. In *Conference companion on Human factors in computing systems - CHI '94* (pp. 413–414). New York, New York, USA: ACM Press. doi:10.1145/259963.260531
Nielsen, J. (1995). Card Sorting to Discover the Users' Model of the Information Space. *Useit*. Retrieved from www.useit.com/papers/sun/cardsort.html
Nielsen, J. (1997). The use and misuse of focus groups. *IEEE Software*, *14*(1), 94–95. doi:10.1109/52.566434
Nielsen, J. (2006). Participation Inequality: Encouraging More Users to Contribute. *Nielsen Norman Group*. Retrieved September 02, 2014, from http://www.nngroup.com/articles/participation-inequality/
Nielsen, J., & Levy, J. (1994). Measuring usability: preference vs. performance. *Communications of the ACM*. doi:10.1145/175276.175282
Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems Empowering people - CHI '90* (pp. 249–256). New York, New York, USA: ACM Press. doi:10.1145/97243.97281
Nielsen, J., & Phillips, V. L. (1993). Estimating the relative usability of two interfaces: heuristic, formal, and empirical methods compared. In S. Ashlund, A. Henderson, E. Hollnagel, K. Mullet, & T. White (Eds.), *Proceedings of the INTERCHI '93 conference on Human factors in computing systems (INTERCHI '93)* (pp. 214–221). Amsterdam, The Netherlands: IOS Press.
Nietzsche, F. (1935). *La Volonté de puissance, tome I*. Gallimard.
Niitamo, V.-P., Kulkki, S., Eriksson, M., & Hribernik, K. A. (2006). State-of-the-Art and Good Practice in the Field of LivingLabs. In *Proceedings of the 12th International Conference on Concurrent Enterprising: Innovative Products and Services through Collaborative Networks (ICE'2006)* (pp. 349–357). Milan.
Ninio, J. (1998). *La science des illusions*. Paris: Odile Jacob.
Norman, D. (1988). *The psychology of everyday things*. New York: Basic Books.
Norman, D. (2004a). *Emotional design: Why we love (or hate) everyday things*. New York: Basic Books.
Norman, D. (2004b). Introduction to This Special Section on Beauty, Goodness, and Usability. *Human-Computer Interaction*, *19*(4), 311–318. doi:10.1207/s15327051hci1904_1
Norman, D. A. (1983a). Design Principles for Human–Computer Interfaces. In *Proc. CHI 1983* (pp. 1–10). ACM Press.
Norman, D. A. (1983b). Steps toward a Cognitive Engineering: Design Rules Based on Analyses of Human Error. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 378–382). ACM Press.
Norman, D. A., Miller, J., & Henderson, A. (1995). What you see, some of what's in the future, and how we go about doing it. In *Conference companion on Human factors in computing systems - CHI '95* (p. 155). New York, New York, USA: ACM Press. doi:10.1145/223355.223477
Nourbakhsh, N., Wang, Y., & Chen, F. (2013). GSR and Blink Features for Cognitive Load Classification. In P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, & M. Winckler (Eds.), *Human-Computer Interaction – INTERACT 2013* (Vol. 8117, pp. 159–166). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-40483-2_11
Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1–18.
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory (3rd ed.)*. New York: McGraw Hill.
O'Brien, H. L., & Toms, E. G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, *59*(6), 938–955. doi:10.1002/asi.20801

O'Reilly, T. (2005). What is Web 2.0: Design Patterns and Business Models for the next generation of software. *O'Reilly Media Inc.* Retrieved August 01, 2004, from http://oreilly.com/web2/archive/what-is-web-20.html

O'Reilly, T. (2007). What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *International Journal of Digital Economics*, *65*, 17–37.

Obrist, M., Geerts, D., Brandtzæg, P. B., & Tscheligi, M. (2008). Design for creating, uploading and sharing user generated content. *Proceeding of the Twenty-Sixth Annual CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI '08*, 2391. doi:10.1145/1358628.1358692

Obrist, M., Law, E. L.-C., Väänänen-Vainio-Mattila, K., Roto, V., Vermeeren, A., & Kuutti, K. (2011). UX research: What Theoretical Roots Do We Build On – If Any? In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11* (p. 165). New York, New York, USA: ACM Press. doi:10.1145/1979742.1979526

Obrist, M., Roto, V., Law, E.-C., Väänänen-Vainio-Mattila, K., Vermeeren, A., & Buie, E. (2012). Theories behind UX research and how they are used in practice. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems (CHI EA '12).* (pp. 2751–2754). New York, USA: ACM. doi:10.1145/2212776.2212712

Obrist, M., Roto, V., & Väänänen-Vainio-Mattila, K. (2009). User experience evaluation: do you know which method to use? In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems - CHI EA '09* (p. 2763). New York, New York, USA: ACM Press. doi:10.1145/1520340.1520401

Obrist, M., Roto, V., Vermeeren, A., Väänänen-Vainio-Mattila, K., Law, E. L.-C., & Kuutti, K. (2012). In search of theoretical foundations for UX research and practice. In *Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts - CHI EA '12* (p. 1979). New York, New York, USA: ACM Press. doi:10.1145/2212776.2223739

Obrist, M., Wright, P. C., Kuutti, K., Rogers, Y., Höök, K., Pyla, P. S., & Frechin, J.-L. (2013). Theory and practice in ux research: uneasy bedfellows? In *CHI '13 Extended Abstracts on Human Factors in Computing Systems (CHI EA '13)* (pp. 2433–2438). New York, USA: ACM. doi:10.1145/2468356.2468795

Ocnarescu, I., Rodio, F., Eve, A., Labrune, J.-B., & Bouchard, C. (2011). Beyond TechCards: a first step toward the investigation of new dimensions of intermediate representations to support the creative process of emerging technologies. In *Proceedings of Design Research Conference by the International Association of Societies of Design Research (IASDR 2011)*.

Olsen, D. (2015). Measure Your Key Metrics. In *The Lean Product Playbook* (pp. 229–258). doi:10.1002/9781119154822.ch13

Olsen, W. (2004). Triangulation in social research : qualitative and quantitative Methods can really be mixed. In M. Holborn (Ed.), *Developments in Sociology*. Ormskirk: Causeway Press.

Olson, G. M., & Moran, T. P. (1998). Commentary on "Damaged Merchandise?." *Human-Computer Interaction*, *13*(3), 263–323. Retrieved from http://portal.acm.org/citation.cfm?id=1462992

Olson, J. R., & Olson, G. (1990). The Growth of Cognitive Modeling in Human-Computer Interaction Since GOMS. *Human-Computer Interaction*, *5*(2), 221–265. doi:10.1207/s15327051hci0502&3_4

Opperud, A. (2004). Semiotic Product Analysis. In D. Hekkert, J. van Erp, & D. Gyi (Eds.), *Design and Emotion: The Experience of Everyday Things* (pp. 137–142). New York: Taylor and Francis.

Ortiz De Guinea, A., Titah, R., & Léger, P. M. (2013). Measure for Measure: A two study multi-trait multi-method investigation of construct validity in IS research. *Computers in Human Behavior*, *29*(3), 833–844. doi:10.1016/j.chb.2012.12.009

Osborn, A. F. (1953). *Applied imagination : Principles and procedures of creative thinking*. New York: Charles Scribner's Sons.

Otondo, R., Vanscotter, J., Allen, D., & Palvia, P. (2007). The complexity of richness: Media, message, and communication outcomes. *Information & Management*, *45*, 21–30. doi:10.1016/j.im.2007.09.003

Pace, S. (2003). *Flow experiences of Web users*. Australian National University.

Pace, S. (2004). A grounded theory of the flow experiences of Web users. *International Journal of Human-Computer Studies*, *60*(3), 327–363. doi:10.1016/j.ijhcs.2003.08.005

Page, S. E. (2007). *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton, NJ: Princeton University Press.

Pahl, G., Wallace, K., & Blessing, L. (2007). *Engineering design : a systematic approach*. London: Springer.

Pak, R., Price, M. M., & Thatcher, J. (2009). Age-sensitive design of online health information : Comparative usability study. *Journal of Medical Internet Research*, *11*(4). doi:10.2196/jmir.1220

Pan, S. J., Ni, X., Sun, J.-T., Yang, Q., & Chen, Z. (2010). Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web - WWW '10* (p. 751). New York, New York, USA: ACM Press. doi:10.1145/1772690.1772767

Pantic, M., & Rothkrantz, L. J. M. (2000). Automatic analysis of facial expressions : the state of the art. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *22*(12), 1424–1445.
Papillo, J. F., & Shapiro, D. (1990). The cardiovascular system. In L. G. Tassinary (Ed.) (Ed.), *Principles of Psychophysiology: Physical, Social, and Inferential Elements* (pp. 456–512). Cambridge: Cambridge University Press.
Partala, T., Surakka, V., & Vanhala, T. (2006). Real-time estimation of emotional experiences from facial expressions. *Interacting with Computers*, *18*(2), 208–226. doi:10.1016/j.intcom.2005.05.002
Partalaa, T., & Surakka, E. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*, *59*(1-2), 185–198. doi:10.1016/S1071-5819(03)00017-X
Pashler, H., Johnstone, J. C., & Ruthruff, E. (2001). Attention and performance. *Annual Review of Psychology*, *52*, 629–651.
Patry, J.-L. (2006). Issues in Critical Multiplism in Evaluation Research: Multiplism of Theories and Analysis of Biases. *Salzburger Beiträge Zur Erziehungswissenschaft*, *10*(1), 23–36.
Paulmann, S., Jessen, S., & Kotz, S. a. (2009). Investigating the Multimodal Nature of Human Communication. *Journal of Psychophysiology*, *23*(2), 63–76. doi:10.1027/0269-8803.23.2.63
Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. doi:10.1080/14786440109462720
Pedhazur, E. J. (1982). *Multiple regression in behavioral research (2nd ed.)*. New York: Holt, Rinehart and Winston.
Peifer, C., Schulz, A., Schächinger, H., Baumann, N., & Antoni, C. H. (2014). The relation of flow-experience and physiological arousal under stress - Can u shape it? *Journal of Experimental Social Psychology*, *53*, 62–69. doi:10.1016/j.jesp.2014.01.009
Pellerey, M. (1996). *Questionario sulle strategie d'apprendimento (QSA)*. Rome, Italy: LAS.
Penichet, V. M. R., Marin, I., Gallud, J. A., Lozano, M. D., & Tesoriero, R. (2007). A Classification Method for CSCW Systems. *Electronic Notes in Theoretical Computer Science (ENTCS)*, *168*(Février), 237–247. doi:DOI= http://dx.doi.org/10.1016/j.entcs.2006.12.007
Pennock, D. M., & Lawrence, S. (2001). The real power of artificial markets. *Science*, *291*, 987–988.
Pennock, D. M., Lawrence, S., Nielsen, F. Å., & Giles, C. L. (2001). Extracting collective probabilistic forecasts from web games. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '01* (pp. 174–183). San Francisco, CA. doi:10.1145/502512.502537
Peter, C., & Herbon, a. (2006). Emotion representation and physiology assignments in digital systems. *Interacting with Computers*, *18*(2), 139–170. doi:10.1016/j.intcom.2005.10.006
Pettigrew, K. E., & McKechnie, L. (E. F. . (2001). The use of theory in information science research. *Journal of the American Society for Information Science and Technology*, *52*(1), 62–73. doi:10.1002/1532-2890(2000)52:1<62::AID-ASI1061>3.0.CO;2-J
Piaget, J. (1967). *Logique et connaissance scientifique*. *Paris: Gallimard*. Paris: Gallimard.
Piaget, J. (1970). *Psychologie et épistémologie*. Paris: Denoel-Gonthier.
Piat, G. (2005). From User-Oriented Design to User-Oriented Technological Design. In *14th International Conference on Management of Technology*. Vienna.
Picard, R., & Scheirer, J. (2001). The Galvactivator: A glove that senses and communicates skin conductivity. In *Proceedings of the 9th International Conference on Human–Computer Interaction*. New Orleans.
Picard, R. W. (1997). *Affective computing*. Cambridge, MA: The MIT Press.
Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(10), 1175–1191. doi:http://dx.doi.org/10.1109/34.954607
Pine, B. J., & Gilmore, J. H. (1999). *The experience economy*. Boston: Harvard Business School Press.
Pinelle, D., & Gutwin, C. (2000). A Review of Groupware Evaluations. In *WETICE '00: Proceedings of the 9th IEEE International Workshops on Enabling Technologies* (pp. 86–91). Washington, DC, USA: IEEE Computer Society.
Pinelle, D., & Gutwin, C. (2002). Groupware walkthrough: adding context to groupware usability evaluation. In *Proceedings of the SIGCHI conference on Human factors in computing systems Changing our world, changing ourselves - CHI '02* (pp. 455–462). New York, New York, USA: ACM Press. doi:10.1145/503376.503458
Pinelle, D., & Gutwin, C. (2008). Evaluating teamwork support in tabletop groupware applications using collaboration usability analysis. *Personal and Ubiquitous Computing*, *12*(3), 237–254. doi:10.1007/s00779-007-0145-4
Pinelle, D., Gutwin, C., & Greenberg, S. (2003). Task analysis for groupware usability evaluation: Modeling shared-workspace tasks with the mechanics of collaboration. *ACM Transactions on Computer-Human Interaction*, *10*(4), 281–311. Retrieved from http://portal.acm.org/citation.cfm?id=966932

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*, *47*(3), 560–569. doi:10.1111/j.1469-8986.2009.00947.x
Pivec, P., & Pivec, M. (2009). Immersed, but How? That Is the Question. *Human IT*, *10*(1), 80–104.
Platon. (1950). *Œuvres complètes (Phèdre), vol. II*. Paris: Gallimard.
Plowman, L., Rogers, Y., & Ramage, M. (1995). What Are Workplace Studies For? In H. Marmolin, Y. Sundblad, & K. Schmidt (Eds.), *Proceedings of the Fourth European Conference on Computer-Supported Cooperative Work ECSCW '95* (pp. 309–324). Dordrecht: Springer Netherlands. doi:10.1007/978-94-011-0349-7_20
Pochon, L.-O. (2008). Créativité et résolution de problèmes. *Résonances*, *7*, 7–9.
Poels, K., Ijsselsteijn, W., & Kort, Y. De. (2008). Development of the Kids Game Experience Questionnaire. In *Meaningful Play Conference*. East Lansing, USA.
Pollard, B. W. (1951). The design, construction, and performance of a large-scale general-purpose digital computer. In *Papers and discussions presented at the Dec. 10-12, 1951, joint AIEE-IRE computer conference: Review of electronic digital computers on - AIEE-IRE '51* (pp. 62–70). New York, New York, USA: ACM Press. doi:10.1145/1434770.1434780
Polson, P. G., Lewis, C., Rieman, J., & Wharton, C. (1992). Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *International Journal of Man-Machine Studies*, *36*(5), 741–773. doi:10.1016/0020-7373(92)90039-N
Poltrock, S., & Grudin, J. (1998). Computer supported cooperative work and groupware. Tutorial notes. In *CH'98 Conference on Human Factors in Computing Systems*.
Popper, K. (1934). *Logik der Forschung*. Mohr Siebeck.
Popper, K. (1985). *Conjectures et réfutations*. Payot.
Popper, K. (1990). Vers une théorie évolutionniste de la connaissance. In *Un univers de propensions*. Éditions de l'Eclat.
Pour, P. A., & Calvo, R. A. (2011). Towards a generic framework for automatic measurements of web usability using affective computing techniques. In S. D'Mello, A. Graesser, B. Schuller, & J.-C. Martin (Eds.), *Proceedings of the 4th international conference on Affective computing and intelligent interaction (ACII'11)* (pp. 447–456). Berlin, Heidelberg: Springer-Verlag.
Prahalad, C. K., & Hamel, G. (1990). The core competence of the corporation. *Harvard Business Review*, *90*(3), 79–91.
Preece, J., Rogers, Y., & Sharp, H. (2002). *Interaction design: beyond human-computer interaction*. New York: John Wiley & Sons.
Preskill, H., & Russ-Eft, D. (2005). *Building evaluation capacity: 72 activities for teaching and training*. Thousand Oaks: Sage.
Priedhorsky, R., Chen, J., Lam, S., Panciera, K., Terveen, L., & Riedl, J. (2007). Creating, destroying, and restoring value in Wikipedia. In *GROUP 2007* (pp. 259–268). ACM Press.
Procci, K., Singer, A. R., Levy, K. R., & Bowers, C. (2012). Measuring the flow experience of gamers: An evaluation of the DFS-2. *Computers in Human Behavior*, *28*(6), 2306–2312. doi:10.1016/j.chb.2012.06.039
Prom Tep, S., Dufresne, A., Saulnier, J., & Archambault, M. (2008). Web site quality evaluation combining eyetracking and physiologicial measures to self-reported emotions: an exploratory research. In *Measuring Behavior'2008* (Vol. 2008, pp. 224–225). Maestricht, Netherland.
Pruitt, J., & Adlin, T. (2006). *The Persona Lifecycle : Keeping People in Mind Throughout Product Design*. Morgan Kaufmann.
Pu, P., Chen, L., & Hu, R. (2012). Evaluating recommender systems from the user's perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction*, *22*(4-5), 317–355. doi:10.1007/s11257-011-9115-7
Qin, H., Patrick Rau, P.-L., & Salvendy, G. (2009). Measuring Player Immersion in the Computer Game Narrative. *International Journal of Human-Computer Interaction*, *25*(2), 107–133. doi:10.1080/10447310802546732
Quinn, A. J., & Bederson, B. B. (2011). Human computation : A Survey and Taxonomy of a Growing Field. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11* (pp. 1403–1412). New York, New York, USA: ACM Press. doi:10.1145/1978942.1979148
Raban, D. R., Moldovan, M., & Jones, Q. (2010). An empirical study of critical mass and online community survival. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work - CSCW '10* (p. 71). New York, New York, USA: ACM Press. doi:10.1145/1718918.1718932
Ramachandran, V. S. (2011). *Le cerveau fait de l'esprit : enquête sur les neurones miroirs*. Dunod.
Ramachandran, V. S., & Hirstein, W. (1997). Three laws of qualia: What neurology tells us about the biological functions of consciousness. *Journal of Consciousness Studies*, *4*(5-6), 429–457.

Rasmussen, J., & Jensen, A. (1974). Mental procedures in real-life tasks: a case study of electronic trouble shooting. *Ergonomics*, *17*(3), 293–307. doi:10.1080/00140137408931355
Rau, P.-L. P., Peng, S.-Y., & Yang, C.-C. (2006). Time distortion for expert and novice online game players. *Cyberpsychology & Behavior : The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, *9*(4), 396–403. doi:10.1089/cpb.2006.9.396
Ravaja, N. (2004). Contributions of psychophysiology to media research : Review and recommendations. *Media Psychology*, *6*, 193–235.
Ravaja, N. (2008). Psychophysiology of digital game playing: The relationship of self-reported emotions with phasic physiological responses. In A. J. Spink, M. R. Ballintijn, N. D. Bogers, F. Grieco, L. W. S. Loijens, L. P. J. J. Noldus, … P. H. Zimmerman (Eds.), *Proceedings of Measuring Behavior*. Maastricht, The Netherlands: Noldus. Retrieved from http://www.noldus.com/mb2008/individual_papers/Symposium vandenHoogen/Symposium_vandenHoogen_Ravaja.pdf
Ravaja, N., Saari, T., Kallinen, K., & Laarni, J. (2006). The Role of Mood in the Processing of Media Messages From a Small Screen: Effects on Subjective and Physiological Responses. *Media Psychology*, *8*(3), 239–265. doi:10.1207/s1532785xmep0803_3
Ravaja, N., Saari, T., Salminen, M., Laarni, J., & Kallinen, K. (2006). Phasic Emotional Reactions to Video Game Events: A Psychophysiological Investigation. *Media Psychology*, *8*(4), 343–367. doi:10.1207/s1532785xmep0804_2
Ray, W. J., & Cole, H. W. (1985). EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes. *Science*, *228*(4700), 750–752. doi:10.1126/science.3992243
Read, J. C., & MacFarlane, S. (2006). Using the fun toolkit and other survey methods to gather opinions in child computer interaction. In *Proceeding of the 2006 conference on Interaction design and children (IDC 06)* (pp. 81–86). New York: ACM Press. doi:10.1145/1139073.1139096
Reason, J. T. (2008). *The human contribution: unsafe acts, accidents and heroic recoveries*. Farnham, England: Ashgate.
Rebolledo-mendez, G., Dunwell, I., & Martínez-mirón, E. A. (2009). Assessing NeuroSky's Usability to Detect Attention Levels in an Assessment Exercise. In J. A. Jacko (Ed.), *Human-Computer Interaction. New Trends* (Vol. 5610, pp. 149–158). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-02574-7
Reeves, B., Lombard, M., & Melwani, G. (1992). Faces on the screen : Pictures or natural experience? In *Annual meeting of the International Communication Association*. Miami, FL.
Reid, S. E., & de Brentani, U. (2004). The Fuzzy Front End of New Product Development for Discontinuous Innovations : A Theoretical Model. *Product Innovation Management*, *21*(3), 170–184.
Reimann, M., Zaichkowsky, J., Neuhaus, C., Bender, T., & Weber, B. (2010). Aesthetic package design: A behavioral, neural, and psychological investigation. *Journal of Consumer Psychology*, *20*(4), 431–441. doi:10.1016/j.jcps.2010.06.009
Remus, U., & Wiener, M. (2010). A multi-method, holistic strategy for researching critical success factors in IT projects. *Information Systems Journal*, *20*(1), 25–52. doi:10.1111/j.1365-2575.2008.00324.x
Renan, E. (1890). *L'Avenir de la science*. Calmann-Levy.
Rens, J.-G. (1984). Révolutions dans la communication: de l'écriture à la télématique. *Sociologie et Sociétés*, *16*(1), 13–22.
Rettig, M. (1994). Prototyping for tiny fingers. *Communications of the ACM*, *37*(4), 21–27. doi:10.1145/175276.175288
Reynolds, T. J., & Gutman, J. (1988). Laddering theory, method, analysis, and interpretation. *Journal of Advertising Research*, *28*(1), 11–31.
Richard Dawkins. (1989). *L'Horloger Aveugle* (Laffont.).
Richter, M., Friedrich, A., & Gendolla, G. H. E. (2008). Task difficulty effects on cardiac activity. *Psychophysiology*, *45*(5), 869–875. doi:10.1111/j.1469-8986.2008.00688.x
Rieder, B., & Röhle, T. (2012). Digital methods: Five challenges. In D. M. Berry (Ed.), *Understanding digital humanities* (pp. 67–84). Basingstoke, England: Palgrave Macmillan.
Riemer, K., Steinfield, C., & Vogel, D. (2009). eCollaboration: On the nature and emergence of communication and collaboration technologies. *Electronic Markets*, *19*(4), 181–188. doi:10.1007/s12525-009-0023-1
Rifkin, J. (2013). *The Third Industrial Revolution: How Lateral Power Is Transforming Energy, the Economy, and the World*. New York: Palgrave Macmillan.
Rifkin, J. (2014). *The zero marginal cost society: The internet of things, the collaborative commons, and the eclipse of capitalism.* New York: Palgrave Macmillan.
Riva, G., Waterworth, J. A., & Waterworth, E. L. (2004). The layers of presence: a bio-cultural approach to understanding presence in natural and mediated environments. *Cyberpsychology & Behavior : The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, *7*(4), 402–16. doi:10.1089/cpb.2004.7.402

Robert, J., & Lesage, A. (2011). From Usability to User Experience with Interactive Systems. In G. A. Boy (Ed.), *The Handbook of Human-Machine Interaction: a Human-Centered Design Approach*. Farnham, Surrey, England: Ashgate.

Robert, J. M. (2003). Que faut-il savoir sur l'utilisateur pour concevoir des interfaces de qualité ? In G. A. Boy (Ed.), *Ingénierie cognitive : IHM et Cognition* (pp. 249–284). Paris: Hermès.

Robert, J.-M. (2008). Vers la plénitude de l'expérience utilisateur. In *Proceedings of the 20th International Conference of the Association Francophone d'Interaction Homme-Machine - IHM '08* (pp. 3–10). New York, New York, USA: ACM Press. doi:10.1145/1512714.1512716

Roberts, B., Harris, M. G., & Yates, T. A. (2005). The roles of inducer size and distance in the Ebbinghaus illusion (Titchener circles). *Perception*, *34*(7), 847–856.

Roberts, T. L., & Moran, T. P. (1982). Evaluation of text editors. In *Proceedings of the 1982 conference on Human factors in computing systems - CHI '82* (pp. 136–141). New York, New York, USA: ACM Press. doi:10.1145/800049.801770

Roberts, T. L., & Moran, T. P. (1983). The evaluation of text editors: methodology and empirical results. *Communications of the ACM*, *26*(4), 265–283. doi:10.1145/2163.2164

Robertson, S. (2001). Requirements trawling: techniques for discovering requirements. *International Journal of Human-Computer Studies*, *55*(4), 405–421. doi:10.1006/ijhc.2001.0481

Roche, S. M., & McConkey, K. M. (1990). Absorption: Nature, assessment, and correlates. *Journal of Personality and Social Psychology*, *59*(1), 91–101. doi:10.1037/0022-3514.59.1.91

Rodio, F., & Bastien, J. M. C. (2013). Heuristics for Video Games Evaluation : How Players Rate Their Relevance for Different Game Genres According to Their Experience. In *Proceedings of the 25th conference francophone on l'Interaction Homme-Machine - IHM '13* (pp. 89–93). doi:10.1145/2534903.2534915

Rodio, F., & Gonguet, A. (2011). Effect of interface type and presentation style on auditor's user experience : a multimodal assessment carried out in a video-conference context. In *Paper session presented at the 2nd Bell Labs Science Workshop*. Villarceaux.

Rodríguez-Sánchez, A. M., Schaufeli, W. B., Salanova, M., & Cifre, E. (2008). Flow experience among information and communication technology users. *Psychological Reports*, *102*(1), 29–39. doi:10.2466/pr0.102.1.29-39

Rogers, R. (2015). Digital Methods for Web Research. In R. A. Scott & S. M. Kosslyn (Eds.), *Emerging Trends in the Behavioral and Social Sciences*. Hoboken, NJ: Wiley.

Root, R. W., & Draper, S. (1983). Questionnaires as a software evaluation tool. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI '83* (pp. 83–87). New York, New York, USA: ACM Press. doi:10.1145/800045.801586

Roseman, I. J., Antoniou, A. A., & Jose, P. E. (1996). Appraisal Determinants of Emotions: Constructing a More Accurate and Comprehensive Theory. *Cognition & Emotion*, *10*(3), 241–278. doi:10.1080/026999396380240

Rosenberg, M. J., & Hovland, C. I. (1960). Cognitive, Affective and Behavioral Components of Attitude. In M. J. Rosenberg, C. I. Hovland, W. J. McGuire, R. P. Abelson, & J. W. Brehm (Eds.), *Attitude organization and change* (pp. 1–14). New Haven: Yale University Press.

Ross, S., Ramage, M., & Rogers, Y. (1995). PETRA: participatory evaluation through redesign and analysis. *Interacting with Computers*, *7*(4), 335–360. doi:10.1016/0953-5438(96)87697-1

Rosson, M. B., & Carroll, J. M. (2001). *Usability Engineering: Scenario-Based Development of Human Computer Interaction*. Massachusetts: Morgan Kaufmann.

Roto, V. (2007). User Experience from Product Creation Perspective. In *Towards a UX Manifesto workshop (COST294- MAUSE), ien conjonction avec HCI 2007* (pp. 31–34). Lancaster, UK.

Roto, V., & Hoonhout, J. (2009). Experiential Evaluation Methods. In *Workshop in the 4th International Conference on Designing Pleasurable Products and Interactions, DPPI 2009*.

Roto, V., Law, E., Vermeeren, A., & Hoonholt, J. (2011). User Experience White Paper. *Dagstuhl Seminar on Demarcating User Experience*. Finland. Retrieved from http://www.allaboutux.org

Roto, V., Law, E., Vermeeren, A., & Hoonhout, J. (2010). Demarcating User eXperience. In *Dagstuhl Seminar*.

Roto, V., & Väänänen-vainio-mattila, K. (2009). User Experience Evaluation Methods in Product Development (UXEM'09). In T. Gross, J. Gulliksen, P. Kotzé, L. Oestreicher, P. Palanque, R. O. Prates, & M. Winckler (Eds.), *Human-Computer Interaction – INTERACT 2009* (Vol. 5727, pp. 981–982). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-03658-3

Roto, V., Vermeeren, A., Väänänen-vainio-mattila, K., & Law, E. (2011). User Experience Evaluation – Which Method to Choose? In P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, & M. Winckler (Eds.), *Human-Computer Interaction – INTERACT 2011* (Vol. 6949, pp. 714–715). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-23768-3

Roto, V., Vermeeren, A., Väänänen-Vainio-Mattila, K., Law, E. L.-C., & Obrist, M. (2012). User Experience Evaluation Methods – Which Method to Choose? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*.

Roto, V., Vermeeren, A., Väänänen-Vainio-Mattila, K., Law, E. L.-C., & Obrist, M. (2013). User Experience Evaluation Methods – Which Method to Choose? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*.

Rowe, D. W., Sibert, J., & Irwin, D. (1998). Heart rate variability: indicator of user state as an aid to human-computer interaction. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '98* (pp. 480–487). New York, New York, USA: ACM Press. doi:10.1145/274644.274709

Rowley, D. E., & Rhoades, D. G. (1992). The cognitive jogthrough : a fast-paced user interface evaluation procedure. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '92* (pp. 389–395). New York, New York, USA: ACM Press. doi:10.1145/142750.142869

Roy, S., Pattnaik, P. K., & Mall, R. (2014). A quantitative approach to evaluate usability of academic websites based on human perception. *Egyptian Informatics Journal*, *15*(3), 159–167. doi:10.1016/j.eij.2014.08.002

Royce, W. W. (1970). Managing the development of large software systems. In *proceedings of IEEE WESCON (Vol. 26, No. 8)*.

Ruhleder, K., & Jordan, B. (1998). Video-Based Interaction Analysis (VBIA) in Distributed Settings: A Tool for Analyzing Multiple-Site, Technology-Supported Interactions. In *Proceedings of the Participatory Design conference (PDC 98)* (pp. 195–196). Seattle WA.

Russell, J. a. (2009). Emotion, core affect, and psychological construction. *Cognition & Emotion*, *23*(7), 1259–1283. doi:10.1080/02699930902809375

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, *57*(3), 493–502. doi:10.1037/0022-3514.57.3.493

Ryan, R. M. (1982). Control and Information in the Intrapersonal Sphere. *Journal of Personality and Social Psychology*, *43*(3), 450–461. doi:10.1037//0022-3514.43.3.450

Ryu, Y. S. (2009). Mobile Phone Usability Questionnaire (MPUQ) and Automated Usability Evaluation. In J. A. Jacko (Ed.), *Human-Computer Interaction. New Trends* (Vol. 5610, pp. 349–351). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-02574-7

Saadé, R., & Bahli, B. (2005). The impact of cognitive absorption on perceived usefulness and perceived ease of use in on-line learning: an extension of the technology acceptance model. *Information & Management*, *42*(2), 317–327. doi:10.1016/j.im.2003.12.013

Saisana, M., Saltelli, A., & Tarantola, S. (2005). Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *168*(2), 307–323. doi:10.1111/j.1467-985X.2005.00350.x

Salvucci, D. (1999). *Mapping eye movements to cognitive processes (Unpublished Doctoral dissertation)*. Carnegie Mellon University.

Sammler, D., Grigutsch, M., Fritz, T., & Koelsch, S. (2007). Music and emotion: electrophysiological correlates of the processing of pleasant and unpleasant music. *Psychophysiology*, *44*(2), 293–304. doi:10.1111/j.1469-8986.2007.00497.x

Sanders, T., & Cairns, P. (2010). Time perception, immersion and music in videogames. In *Proceedings of the 24th BCS Interaction Specialist ...* (pp. 160–167). Swinton, UK, UK: British Computer Society. Retrieved from http://dl.acm.org/citation.cfm?id=2146327

Sandri, E. (2013). La sérendipité sur Internet : égarement documentaire ou recherche créatrice? »,. *Cygne Noir : Revue D'exploration Sémiotique*, *1*.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web (WWW '01)* (pp. 285–295). New York, NY, USA: ACM. doi:10.1145/371920.372071

Sauro, J. (2008). Task Times In Formative Usability Tests. *measuring U*. Retrieved from http://www.measuringu.com/formative-time.php

Sauro, J. (2011a). *A practical guide to the System Usability Scale (SUS): Background, benchmarks & best practices*. Denver, CO: Measuring Usability LLC.

Sauro, J. (2011b). The Standardized Universal Percentile Rank Questionnaire (SUPR-Q). Retrieved April 02, 2016, from www .suprq.com/

Sauro, J. (2014). The relationship between problem frequency and problem severity in usability evaluations. *Journal of Usability Studies*, *10*(1), 17–25.

Sauro, J. (2015). SUPR-Q: a comprehensive measure of the quality of the website user experience. *Journal of Usability Studies*, *10*(2), 68–86.

Sauro, J., & Dumas, J. S. (2009). Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI '09* (p. 1599). New York, New York, USA: ACM Press. doi:10.1145/1518701.1518946

Sauro, J., & Kindlund, E. (2005). A method to standardize usability metrics into a single score. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '05* (p. 401). New York, New York, USA: ACM Press. doi:10.1145/1054972.1055028

Sauro, J., & Kindlund, E. (2005). How long should a task take? identifying specification limits for task times in usability tests. In *Proceeding of the Human Computer Interaction International Conference (HCII 2005)*. Las Vegas.

Sauro, J., & Lewis, J. R. (2005). Estimating Completion Rates from Small Samples using Binomial Confidence Intervals: Comparisons and Recommendations. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting (HFES 2005)*. Orlando, FL, USA.

Sauro, J., & Lewis, J. R. (2009). Correlations among prototypical usability metrics: Evidence for the Construct of Usability. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI '09* (p. 1609). New York, New York, USA: ACM Press. doi:10.1145/1518701.1518947

Sauro, J., & Lewis, J. R. (2010). Average task times in usability tests: what to report? In *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10* (p. 2347). New York, New York, USA: ACM Press. doi:10.1145/1753326.1753679

Sauro, J., & Lewis, J. R. (2012). *Quantifying the User Experience: Practical Statistics for User Research*. Morgan Kaufmann.

Sauteron, F. (2009). *La chute de l'empire Kodak*. Paris: Harmattan.

Scapin, D. L. (1988). *Vers des outils formels de description des tâches orientés conception d'interfaces (Rapport de recherche N° 893)*. Rocquencourt, France.

Scapin, D. L., & Bastien, J. M. C. (2001). Analyse des tâches et aide ergonomique à la conception : l'approche MAD*. In C. Kolski (Ed.), *Systèmes d'information et Interactions homme-machine*. Paris: Hermés.

Schaffer, N. (2009). *Verifying an Integrated Model of Usability in Game*. Rensselaer Polytechnic Institute.

Scherer, K. R. (1989). Vocal measurement of emotion. In R. Plutchik & H. Kellerman (Eds.) (Eds.), *Emotion: Theory, research, and experience (Vol. 4)* (pp. 233–239). San Diego, CA: Academic Press, Inc.

Scherer, K. R. (2001). Appraisal considered as a process of multi-level sequential checking. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, Methods, Research* (pp. 92–120). Oxford: Oxford University Press.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, *44*(4), 695–729. doi:10.1177/0539018405058216

Schmitt, F., Kindsm, M. C., & Herczeg, M. (2010). Mental Models of Disappearing Systems : Challenges for a Better Understanding. In *Proceedings of the Sixth International Workshop on Modeling and Reasoning in Context* (pp. 61–72). Lisbon, Portugal: CEUR-WS.org.

Schmitt, M. (2006). Conceptual, theoretical, and historical foundations of multimethod assessment. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology*. Washington, DC: American Psychological Association.

Schubert, E. (1999). Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. *Australian Journal of Psychology*, *51*(3), 154–165. doi:10.1080/00049539908255353

Schubert, E. (2001). Continuous Measurement of Self-Report Emotional Response to Music. In P. Juslin & J. Sloboda (Eds.), *Music and Emotion: Theory and Research* (pp. 393–414). Oxford University Press.

Schuler, D., & Namioka, A. (1993). *Participatory Design: Principles and Practices*. Hillsdale, NJ: Erlbaum.

Schulze, G. (1993). *Die Erlebnisgesellschaft [The Experience Society]*. Frankfurt: Campus Verlag.

Schumm, J., Setz, C., Bächlin, M., Bächler, M., Arnrich, B., & Tröster, G. (2010). Unobtrusive physiological monitoring in an airplane seat. *Personal and Ubiquitous Computing*, *14*(6), 541–550. doi:10.1007/s00779-009-0272-1

Schutt, R., & O'Neil, C. (2014). *Doing Data Science : straight talk from the frontline*. Sebastopol, CA: O'Reilly Media.

Schütte, R. (2006). *Developing an expert program software for Kansei Engineering*. Linkoping University, Sweden.

Schwartz, G. E., Weinberger, D. a, & Singer, J. a. (1981). Cardiovascular differentiation of happiness, sadness, anger, and fear following imagery and exercise. *Psychosomatic Medicine*, *43*(4), 343–364. doi:10.1097/00006842-198108000-00007

Schwatrz, B. (2004). *The Paradox of Choice*. New York: Ecco.

Schweizer, K. (2011). Some Thoughts Concerning the Recent Shift from Measures with Many Items to Measures with Few Items. *European Journal of Psychological Assessment*, *27*(2), 71–72. doi:10.1027/1015-5759/a000056

Scollon, C. N., Kim-Prieto, C., & Diener, E. (2003). Experience Sampling: Promises and Pitfalls, Strengths and Weaknesses. *Journal of Happiness Studies*, *4*(1), 5–34. doi:10.1023/A:1023605205115

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39–83). Chicago: Rand-McNally.

Scriven, M. (1974). Evaluation perspectives and procedures. In J. W. Popham (Ed.), *Evaluation in education: Current application* (pp. 3–93). Berkeley, CA: McCutcheon.
Scriven, M. (1983). Evaluation ideologies. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 229–260). Boston: Kluwer-Nijhoff.
Scriven, M. (1991). *Evaluation thesaurus (4th ed.)*. Thousand Oaks, CA: Sage.
Seffah, A., Donyaee, M., Kline, R. B., & Padda, H. K. (2006). Usability measurement and metrics: A consolidated model. *Software Quality Journal*, *14*(2), 159–178. doi:10.1007/s11219-006-7600-8
Segall, M., Campbell, D., & Herskovits, M. J. (1966). *The influence of culture on visual perception*. New York: Bobbs-Merrill.
Segall, M. H. (1979). *Cross-Cultural Psychology: Human Behavior in Global Perspective*. Monterey, CA: Brooks/Cole.
Segrestin, B. (2006). *Innovation et coopération interentreprises: comment gérer les partenariats d'exploration?*. Paris: CNRS Éditions.
Seligman, M. E. P. (1999). The President's Address (1998 APA Annual Report). *American Psychologist*, *54*, 559–562.
Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction. *American Psychologist*, *55*(1), 5–14. doi:10.1037//0003-066X.55.1.5
Selvin, H. C., & Stuart, A. (1966). Data-dredging procedures in survey analysis. *The American Statistician*, *20*(3), 20–23.
Serrano, B., Botella, C., Baños, R. M., & Alcañiz, M. (2013). Using virtual reality and mood-induction procedures to test products with consumers of ceramic tiles. *Computers in Human Behavior*, *29*(3), 648–653. doi:10.1016/j.chb.2012.10.024
Serres, M. (1969). *Hermès: la communication*. Paris: Éditions de Minuit.
Serres, M. (2007). Des révolutions cognitives. In *Les 40 ans de l'INRIA*. Lille.
Serres, M., & Bensaude-Vincent, B. (1989). *Éléments d'histoire des sciences*. Bordas.
Sesma, L., Villanueva, A., & Cabeza, R. (2012). Evaluation of pupil center-eye corner vector for gaze estimation using a web cam. In *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12* (Vol. 1, p. 217). New York, New York, USA: ACM Press. doi:10.1145/2168556.2168598
Shackel, B. (1959). Ergonomics for a computer. *Design*, *120*, 36–39.
Shackel, B. (1981). The concept of usability. In *Proceedings of IBM Software and Information Usability Symposium* (pp. 1–30). Poughkeepsie, NY: IBM Corporation.
Shackel, B. (2009a). Designing for people in the age of information. *Interacting with Computers*, *21*(5-6), 325–330. doi:10.1016/j.intcom.2009.04.006
Shackel, B. (2009b). Usability – Context, framework, definition, design and evaluation. *Interacting with Computers*, *21*(5-6), 339–346. doi:10.1016/j.intcom.2009.04.007
Shadish, W. R. (1986). Planned critical multiplism : Some elaborations. *Behavioral Assessment*, *8*, 75–103.
Shan, C., Chen, W., Wang, H., & Song, M. (2015). *The Data Science Handbook : Advice and Insights from 25 Amazing Data Scientists*. Data Science Bookshelf.
Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(4), 623–656. doi:10.1002/j.1538-7305.1948.tb00917.x
Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, *30*(1), 50–64.
Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. University of Illinois Press.
Shastri, D., Pavlidis, I., & Wesley, A. (2009). A Method to Monitor Operator Overloading. In J. A. Jacko (Ed.), *Human-Computer Interaction. New Trends* (Vol. 5610, pp. 169–175). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-02574-7
Shavelson, R. J. (2009). Lee J. cronbach 1916—2001. *National Academy of Sciences Biographical Memoir*. Retrieved from http://www.nasonline.org/publications/biographical-memoirs/memoir-pdfs/Cronbach_Lee_J.pdf
Shedroff, N. (2001). *Experience design 1*. Indianapolis, Ind: New Riders Pub.
Shelton-Rayner, G. K., Mian, R., Chandler, S., Robertson, D., & Macdonald, D. W. (2012). Leukocyte responsiveness, a quantitative assay for subjective mental workload. *International Journal of Industrial Ergonomics*, *42*(1), 25–33. doi:10.1016/j.ergon.2011.11.004
Sherry, J. L. (2004). Flow and Media Enjoyment. *Communication Theory*, *14*(4), 328–347. doi:10.1093/ct/14.4.328
Shneiderman, B. (1980). *Software Psychology: Human Factors in Computer and Information Systems*. Cambridge, MA: Winthrop.
Shneiderman, B. (2011). Technology-Mediated Social Participation: The Next 25 Years of HCI Challenges. In J. A. Jacko (Ed.), *Human-Computer Interaction. Design and Development Approaches* (Vol. 6761, pp. 3–14). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-21602-2

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations : uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.
Siefert, C. J., Kothuri, R., Jacobs, D. B., Levine, B., Plummer, J., & Marci, C. D. (2009). Winning the Super "Buzz" Bowl. *Journal of Advertising Research*, *49*(3), 293. doi:10.2501/S0021849909090424
Siegle, G. J., Ichikawa, N., & Steinhauer, S. (2008). Blink before and after you think: blinks occur prior to and following cognitive load indexed by pupillary responses. *Psychophysiology*, *45*(5), 679–687. doi:10.1111/j.1469-8986.2008.00681.x
Silverman, D. (2005). *Doing qualitative research : a practical handbook*. London: Sage Publications.
Sim, G., & Horton, M. (2012). Investigating Children´s Opinions of Games: Fun Toolkit vs. This or That. In *Proceedings of the 11th International Conference on Interaction Design and Children(IDC '12)* (Vol. 47, pp. 70–77). New York: ACM. doi:10.1145/2307096.2307105
Sime, M. E., Green, T. R. G., & Guest, D. J. (1973). Psychological evaluation of two conditional constructions used in computer languages. *International Journal of Man-Machine Studies*, *5*(1), 105–113.
Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, *69*(1), 99–118.
Simon, H. A. (1959). Theories of Decision-Making in Economics and Behavioral Science. *American Economic Review*, *49*(1), 253–283.
Simon, H. A., & Newell, A. (1963). The uses and limitations of models. In M. Marx (Ed.), *Theories in contemporary psychology* (pp. 89–104). New York: MacMillan.
Simondon, G. (1958). *Du mode d'existence des objets techniques*. Paris: Aubier.
Siné, A., Hausswalt, P., & Garcin, C. (2012). *Le soutien à l'économie numérique et à l'innovation, Rapport n°2011-M-060-02*. Paris, France.
Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments (five): Speculations on the role of presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, *6*, 603–616.
Slaughter, L., Harper, B., & Norman, K. (1994). Assessing the equivalence of the paper and on-line formats of the QUIS 5.5. In *Proceedings of the 2nd Annual Mid-Atlantic Human Factors Conference* (pp. 87–91). Washington: HFES.
Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19.
Snow, C. P. (1993). *The two cultures*. Cambridge: Cambridge University Press.
Snow, R. E. (1986). On intelligence. In R. J. Sternberg & D. K. Detterman (Eds.), *What is intelligence?* (pp. 133–139). Norwood, NJ: Ablex.
Solingen, R., & Berghout, E. (1999). *The goal/question/metric method: A practical guide for quality improvement of software development*. London: McGraw-Hill.
Sonnenwald, D. H., Maglaughlin, K. L., & Whitton, M. C. (2001). Using innovation diffusion theory to guide collaboration technology evaluation: work in progress. In *Proceedings Tenth IEEE International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises. WET ICE 2001* (pp. 114–119). IEEE Comput. Soc. doi:10.1109/ENABL.2001.953399
Spearman, C. (1904). "General intelligence", objectively determined ans measured. *American Journal of Psychology*, *15*, 201–292.
Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*(3), 271–295.
Sperandio, J.-C. (2001). Critères ergonomiques de l'assistance technologiques aux opérateurs. In *JIM'2001 : Interaction Homme - Machine & Assistance*. Metz, France.
Ståhl, A., Höök, K., Svensson, M., Taylor, A. S., & Combetto, M. (2009). Experiencing the affective diary. *Personal and Ubiquitous Computing*, *13*(5), 365–378. doi:10.1007/s00779-008-0202-7
Stern, R. M., Ray, W. J., & Quigley, K. . (2001). *Psychophysiological Recording*. New York: Oxford University Press.
Steves, M. P., Morse, E., Gutwin, C., & Greenberg, S. (2001). A comparison of usage evaluation and inspection methods for assessing groupware usability. In *Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work* (pp. 125–134). ACM. Retrieved from http://portal.acm.org/citation.cfm?id=500306
Steyer, R., & Schmitt, M. (1990). The Effects of Aggregation Across and Within Occasions on Consistency, Specificity, and Reliability. *Methodika*, *4*, 58–94.
Stiegler, B. (1999). L'hyperindustrialisation de la culture et le temps des attrape-nigauds. *Art Press*, *hors série*.
Stiegler, B. (2009). Technologies culturelles et économie de la contribution. *Culture et Recherche*, *121*, 30–31.
Stiegler, B. (2013). Le blues du Net. *Blogs lemonde.fr (Lois des réseaux)*. Retrieved October 23, 2014, from http://reseaux.blog.lemonde.fr/2013/09/29/blues-net-bernard-stiegler/
Stiegler, B., & Ars Industrialis. (2008). *Réenchanter le monde: La valeur esprit contre le populisme industriel*. Paris: Flammarion.
Stiemerling, O., & Cremers, A. B. (1998). The use of cooperation scenarios in the design and evaluation of a CSCW system. *IEEE Transactions on Software Engineering*, *24*(12), 1171–1181. doi:10.1109/32.738345

Stiglitz, J. E., Escoto, B. M., Chemla, F., & Chemla, P. (2010). *Le rapport Stiglitz: Pour une vraie réforme du système monétaire et financier international après la crise mondiale*. Brignon: Les Liens qui libèrent.
Stone, A. A., & Litcher-kelly, L. (2006). Real-world data momentary capture of real-world data. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 61–72). Washington, DC: American Psychological Association.
Stonier, T. (1983). *The Wealth of Information*. London: Methuan.
Suchman, L. (1987). *Plans and Situated Actions: The problem of human-machine communication*. Cambridge, UK: Cambridge University Press.
Suh, E., Diener, E., & Fujita, F. (1996). Events and subjective well-being: Only recent events matter. *Journal of Personality and Social Psychology*, *70*, 1091 – 1102.
Sundar, S. S., Xu, Q., Bellur, S., Oh, J., & Jia, H. (2010). Modality is the message: interactivity effects on perception and engagement. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems - CHI EA '10* (p. 4105). New York, New York, USA: ACM Press. doi:10.1145/1753846.1754110
Surowiecki, J. (2005). *The wisdom of crowds*. New York: Doubleday.
Suzuki, S., Matsui, T., Kagawa, M., Asao, T., & Kotani, K. (2013). An approach to a non-contact vital sign monitoring using dual-frequency microwave radars for elderly care. *Journal of Biomedical Science and Engineering*, *6*, 704–711. doi:10.4236/jbise.2013.67086
Swallow, D., Blythe, M., & Wright, P. (2005). Grounding experience: relating theory and method to evaluate the user experience of smartphones. In *Proceedings of the EACE'05* (pp. 91–98).
Sward, D. (2006). *Gaining a competitive advantage through user experience design*. Retrieved from http://www.intel.com/it/pdf/comp-adv-user-exp.pdf
Szafir, D., & Mutlu, B. (2012). Pay attention! Designing Adaptive Agents that Monitor and Improve User Engagement. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12* (p. 11). New York, New York, USA: ACM Press. doi:10.1145/2207676.2207679
Tähti, M., & Arhippainen, L. (2004). A Proposal of collecting Emotions and Experiences. In *Proceedings of HCI 2004* (p. volume 2. 6–10). Leeds, England.
Taillet, R., Villain, L., & Febvre, P. (2013). *Dictionnaire de physique* (p. 474). Bruxelles: De Boeck.
Tang, A., & Boring, S. (2012). #EpicPlay: Crowd-sourcing Sports Video Highlights. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12* (p. 1569). New York, New York, USA: ACM Press. doi:10.1145/2207676.2208622
Tang, J. C. (1991). Findings from observational studies of collaborative work. *International Journal of Man-Machine Studies*, *34*(2), 143–160. doi:10.1016/0020-7373(91)90039-A
Tanner, P. P., & Buxton, W. A. S. (1985). Some Issues in Future User Interface Management System (UIMS) Development. In G. E. Pfaff (Ed.), *User Interface Management Systems* (pp. 67–79). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-70041-5_8
Tapscott, D., & Williams, A. D. (2006). *Wikinomics: How mass collaboration changes everything*. *Tapscott, D., & Williams, A. D.* New York: Portfolio.
Tashakkori, A., & Teddlie, C. (1998). *Mixed Methodology : Combining Qualitative and Quantitative Approaches*. Thousand Oaks, CA: Sage.
Tedesco, D. P., & Tullis, T. (2006). A Comparison of Methods for Eliciting Post-Task Subjective Ratings in Usability Testing. In *Usability Professionals Association (UPA) Conference*. Denver, Colorado.
Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic Capabilities and Strategic Management. *Strategic Management Journal*, *18*(7), 509–533.
Ter Heerdt, J., & Bondarouk, T. (2009). Information Overload in the New World of Work: Qualitative Study into the Reasons and Countermeasures. In T. Bondarouk, H. Ruel, K. Guiderdoni-Jourdain, & E. Oiry (Eds.), *Handbook of Research on E-Transformation and Human Resources Management Technologies* (pp. 396–418). IGI Global. doi:10.4018/978-1-60566-304-3
Terzis, V., Moridis, C. N., & Economides, A. A. (2010). Measuring instant emotions during a self-assessment test. In *Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research - MB '10* (pp. 1–4). New York, New York, USA: ACM Press. doi:10.1145/1931344.1931362
Tezza, R., Bornia, A. C., & de Andrade, D. F. (2011). Measuring web usability using item response theory: Principles, features and opportunities. *Interacting with Computers*, *23*(2), 167–175. doi:10.1016/j.intcom.2011.02.004
Thiran, J.-P., Bourlard, H., & Marques, F. (2010). *Multimodal signal processing : theory and applications for human-computer interaction*. Amsterdam ; Boston ; London: Elsevier.
Thomke, S. H. (1998). Managing Experimentation in the Design of New Products. *Management Science*, *44*(6), 743–762.
Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286. doi:10.1037/h0070288

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: Univiversity of Chicago Press.
Timo, J. (2005). Satisfied and Dissatisfied at the Same Time: The 'Must-Have' and "Attractive" Properties of a User Interface. In *UIQM'2005 : 1st International Workshop on User Interface Web Quality Models* (pp. 56–61). Rome, Italy.
Toomim, M., Kriplean, T., Pörtner, C., & Landay, J. (2011). Utility of human-computer interactions: toward a science of preference measurement. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11* (p. 2275). New York, New York, USA: ACM Press. doi:10.1145/1978942.1979277
Tractinsky, N. (1997). Aesthetics and apparent usability. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '97* (pp. 115–122). New York, New York, USA: ACM Press. doi:10.1145/258549.258626
Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers*, *13*(2), 127–145.
Tractinsky, N., & Meyer, J. (2001). Task structure and the apparent duration of hierarchical search. *International Journal of Human–Computer Studies*, *55*(5), 845–860.
Traphagan, T. W., Chiang, Y. V., Chang, H. M., Wattanawaha, B., Lee, H., Mayrath, M. C., … Resta, P. E. (2010). Cognitive, social and teaching presence in a virtual world and a text chat. *Computers & Education*, *55*(3), 923–936. doi:10.1016/j.compedu.2010.04.003
Trevino, L. K., & Webster, J. (1992). Flow in computer-mediated communication. *Communication Research*, *19*(1), 539–573.
Treynor, J. L. (1987). Market Efficiency and the Bean Jar Experiment. *Financial Analysts Journal*, *43*(3), 50–53.
Truong, K. N., Hayes, G. R., & Abowd, G. D. (2006). Storyboarding: an empirical determination of best practices and effective guidelines. In *Proceedings of the 6th ACM conference on Designing Interactive systems - DIS '06* (p. 12). New York, New York, USA: ACM Press. doi:10.1145/1142405.1142410
Tuch, A. N., Bargas-Avila, J. a., Opwis, K., & Wilhelm, F. H. (2009). Visual complexity of websites: Effects on users' experience, physiology, performance, and memory. *International Journal of Human-Computer Studies*, *67*(9), 703–715. doi:10.1016/j.ijhcs.2009.04.002
Tullis, T. S., & Albert, B. (2008). *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Amsterdam: Elsevier/Morgan Kaufmann.
Tullis, T. S., & Stetson, J. (2004). A comparison of questionnaires for assessing website usability. In *UPA 2004 conference*. Minneapolis, MN: UPA.
Tullock, G. (2001). A Comment on Daniel Klein's "A Plea to Economists Who Favor Liberty." *Eastern Economic Journal*, *27*(2), 203–207.
Tung, F.-W., & Deng, Y.-S. (2006). Designing social presence in e-learning environments: Testing the effect of interactivity on children. *Interactive Learning Environments*, *14*(3), 251–264. doi:10.1080/10494820600924750
Twidale, M., Randall, D., & Bentley, R. (1994). Situated evaluation for cooperative systems. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work - CSCW '94* (pp. 441–452). New York, New York, USA: ACM Press. doi:10.1145/192844.193066
Ulle, F. (2010). The Psychophysiology of Flow During Piano Playing. *Emotion*, *10*(3), 301–311. doi:10.1037/a0018432
Ullman, D. G. (2010). *The mechanical design process*. Boston: McGraw-Hill Higher Education.
Ulrich, M., Keller, J., Hoenig, K., Waller, C., & Grön, G. (2014). Neural correlates of experimentally induced flow experiences. *NeuroImage*, *86*, 194–202. doi:10.1016/j.neuroimage.2013.08.019
Urban, G. L., & von Hippel, E. (1988). Lead User Analyses for the Development of New Industrial Products. *Management Science*, *34*(5), 569–582. doi:10.1287/mnsc.34.5.569
Urquijo, S. P., Scrivener, S. A. R., & Palmén, H. K. (1993). The use of breakdown analysis in synchronous CSCW system design. In G. de Michelis, C. Simone, & K. Schmidt (Eds.), *Proceedings of the Third European Conference on Computer-Supported Cooperative Work 13–17 September 1993, Milan, Italy ECSCW '93* (pp. 281–293). Dordrecht: Springer Netherlands. doi:10.1007/978-94-011-2094-4_19
Väänänen-Vainio-Mattila, K., Roto, V., & Hassenzahl, M. (2008). Now let's do it in practice: user experience evaluation methods in product development. In *Proceeding of the twenty-sixth annual CHI conference extended abstracts on Human factors in computing systems - CHI '08* (p. 3961). New York, New York, USA: ACM Press. doi:10.1145/1358628.1358967
Väänänen-Vainio-Mattila, K., & Segerståhl, K. (2009). A Tool for Evaluating Service User eXperience (ServUX) : Development of a Modular Questionnaire. In *User Experience Evaluation Methods, UXEM'09 Workshop at Interact (Vol. 9)*.
Väänänen-Vainio-Mattila, K., Väätäjä, H., & Vainio, T. (2009). Opportunities and Challenges of Designing the Service User eXperience (SUX) in Web 2.0. In H. Isomäki & P. Saariluoma (Eds.), *Future Interaction Design II* (pp. 117–139). London: Springer. doi:10.1007/978-1-84800-385-9_6

Vaananen-Vainio-Mattila, K., & Waljas, M. (2010). Evaluating user experience of cross-platform web services with a heuristic evaluation method. *International Journal of Arts and Technology*, *3*(4), 402–421. doi:10.1504/IJART.2010.03583

Väänänen-Vainio-Mattila, K., & Wäljas, M. (2009). Developing an expert evaluation method for user eXperience of cross-platform web services. In A. Lugmayr, H. Franssila, O. Sotamaa, P. Näränen, & J. Vanhala (Eds.), *Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era (MindTrek '09)* (p. 162). New York: ACM. doi:10.1145/1621841.1621871

Väätäjä, H., Koponen, T., & Roto, V. (2009). Developing practical tools for user experience evaluation: a case from mobile news journalism. In L. Norros, H. Koskinen, L. Salo, & P. Savioja (Eds.), *European Conference on Cognitive Ergonomics: Designing beyond the Product - Understanding Activity and User Experience in Ubiquitous Environments (ECCE '09)*. Finland: VTT Technical Research Centre of Finland.

Valery, P. (1959). *Cahiers XIII*. Paris: Ed. du CNRS.

Van Boven, L., & Gilovich, T. (2003). To Do or to Have? That Is the Question. *Journal of Personality and Social Psychology*, *85*(6), 1193–1202.

Van Boxtel, G. J. M., Denissen, A. J. M., Jäger, M., Vernon, D., Dekker, M. K. J., Mihajlović, V., & Sitskoorn, M. M. (2012). A novel self-guided approach to alpha activity training. *International Journal of Psychophysiology*, *83*(3), 282–294. doi:10.1016/j.ijpsycho.2011.11.004

Van den Hoogen, W. M., IJsselsteijn, W. A., & de Kort, Y. A. W. (2009). Effects of Sensory Immersion on Behavioural Indicators of Player Experience : Movement Synchrony and Controller Pressure. In *Breaking New Ground: Innovation in Games, Play, Practice and Theory: Proceedings of the 2009 Digital Games Research Association Conference*.

Van der Veer, G., Lenting, B. F., & Bergevoet, B. A. J. (1996). GTA: Groupware task analysis — Modeling complexity. *Acta Psychologica*, *91*(3), 297–322. doi:10.1016/0001-6918(95)00065-8

Van der Veer, G., & van Welie, M. (2000). Task based groupware design : putting theory into practice. In *Proceedings of the conference on Designing interactive systems processes, practices, methods, and techniques - DIS '00* (pp. 326–337). New York, New York, USA: ACM Press. doi:10.1145/347642.347781

Van schaik, P., & Ling, J. (2008). Modelling user experience with web sites: Usability, hedonic value, beauty and goodness. *Interacting with Computers*, *20*(3), 419–432. doi:10.1016/j.intcom.2008.03.001

Vanden Abeele, V., Zaman, B., & De Grooff, D. (2012). User eXperience Laddering with preschoolers : Unveiling attributes and benefits of cuddly toy interfaces. *Personal and Ubiquitous Computing*, *16*(4), 451–465. doi:10.1007/s00779-011-0408-y

Vanhée, N. (2008). *La coordination des savoirs au sein de partenariats d'innovation*. Université Louis Pasteur, Strasbourg I.

Varenne, F. (2011). *Modéliser le social : Méthodes fondatrices et évolutions récentes*. Paris: Dunod.

Veltman, J. A., & Gaillard, A. W. (1996). Physiological indices of work- load in a simulated flight task. *Biological Psychology*, *42*(3), 323–342.

Veltman, J. A., & Gaillard, A. W. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics*, *41*(5), 656–669.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478. Retrieved from http://www.jstor.org/stable/30036540

Verdot, V., Saidi, A., & Fournigault, L. (2011). Virtual Hybrid Communications – A Telecom Infrastructure for the Metaverse. *Journal of Virtual World Reasearch*, *4*(3).

Vermeeren, A., Kort, J., Cremers, A., Smets, N., & Fokker, J. (2008). Comparing UX measurements : a case study. In E. L.-C. Law, N. Bevan, G. Christou, M. Springett, & M. Lárusdóttir (Eds.), *Proceedings of COST294-workshop "Meaningful measures: Valid Useful User experience Measurement"* (pp. 72–78). Reykjavik, Iceland.

Vermeeren, A. P. O. S., Law, E. L., Roto, V., Obrist, M., Hoonhout, J., & Väänänen-Vainio-Mattila, K. (2010). User experience evaluation methods : current state and development needs. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction Extending Boundaries - NordiCHI '10* (p. 521). New York, New York, USA: ACM Press. doi:10.1145/1868914.1868973

Visch, V. T., Tan, E. S., & Molenaar, D. (2010). The emotional and cognitive effect of immersion in film viewing. *Cognition & Emotion*, *24*(8), 1439–1445. doi:10.1080/02699930903498186

Vizcaino, A., Piattini, M., Martinez, M., & Aranda, G. (2005). Evaluating Collaborative Applications from a Knowledge Management Approach. In *14th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise (WETICE'05)* (pp. 221–225). IEEE. doi:10.1109/WETICE.2005.36

Von Ahn, L. (2005). *Human Computation*. Carnegie Mellon University.

Von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04* (pp. 319–326). New York, New York, USA: ACM Press. doi:10.1145/985692.985733

Von Eye, A., & Mun, E. Y. (2005). *Analyzing rater agreement : manifest variable methods*. (N. Mahwah, Ed.) (Vol. 101). Lawrence Erlbaum. doi:10.1198/jasa.2006.s107

Von Hippel, E. (1986). Lead Users : A Source of Novel Product Concepts. *Management Science*, *32*(7), 791–806. doi:doi:10.1287/mnsc.32.7.791

Von Hippel, E. (2014). Open User Innovation. In M. Soegaard & R. F. Dam (Eds.), *The Encyclopedia of Human-Computer Interaction, 2nd Ed.* Aarhus, Denmark: The Interaction Design Foundation.

Vorderer, P., Klimmt, C., & Ritterfeld, U. (2004). Enjoyment: At the Heart of Media Entertainment. *Communication Theory*, *14*(4), 388–408. doi:10.1111/j.1468-2885.2004.tb00321.x

Voss, K. E., Spangenberg, E. R., & Grohmann, B. (2003). Measuring the Hedonic and Utilitarian Dimensions of Consumer Attitude. *Journal of Marketing Research*, *40*(3), 310–321. doi:10.1509/jmkr.40.3.310.19238

Waitelonis, J., & Sack, H. (2012). Towards exploratory video search using linked data. *Multimedia Tools and Applications*, *59*(2), 645–672. doi:10.1007/s11042-011-0733-1

Wakefield, R., & Whitten, D. (2006). Mobile Computing: A User Study on Hedonic/Utilitarian Mobile Device Usage. *European Journal of Information Systems*, *15*(3), 292–300.

Wakkary, R., & Hatala, M. (2006). ec(h)o : situated play in a tangible and audio museum guide. In *Proceedings of the 6th ACM conference on Designing Interactive systems - DIS '06* (pp. 281–290). New York, NY: ACM Press. doi:10.1145/1142405.1142448

Walker, B. N., & Lindsay, J. (2006). Navigation performance with a virtual auditory display : Effects of beacon sound, capture radius, and practice. *Human Factors*, *48*(2), 265–278.

Wang, H. C., & Chiu, Y. F. (2011). Assessing e-learning 2.0 system success. *Computers & Education*, *57*(2), 1790–1800. doi:10.1016/j.compedu.2011.03.009

Wang, J., & Senecal, S. (2007). Measuring Perceived Website Usability. *Journal of Internet Commerce*, *6*(4), 97–112. doi:10.1080/15332860802086318

Wang, L., Duffy, V. G., & Du, Y. (2007). A Composite Measure for the Evaluation of Mental Workload. In V. G. Duffy (Ed.), *Digital Human Modeling* (Vol. 4561, pp. 460–466). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-73321-8

Wang, N., Kosinski, M., Stillwell, D. J., & Rust, J. (2014). Can Well-Being be Measured Using Facebook Status Updates? Validation of Facebook's Gross National Happiness Index. *Social Indicators Research*, *115*(1), 483–491. doi:10.1007/s11205-012-9996-9

Ward, R. (2004). An analysis of facial movement tracking in ordinary human-computer interaction. *Interacting with Computers*, *16*(5), 879–896. doi:10.1016/j.intcom.2004.08.002

Ward, R. D., & Marsden, P. H. (2003). Physiological responses to different WEB page designs. *International Journal of Human-Computer Studies*, *59*(1-2), 199–212. doi:10.1016/S1071-5819(03)00019-3

Wasserman, A. I., & Shewmake, D. T. (1982). Rapid prototyping of interactive information systems. *ACM SIGSOFT Software Engineering Notes*, *7*(5), 171–180. doi:10.1145/1006258.1006289

Waterink, W., & van Boxtel, A. (1994). Facial and jaw-elevator EMG activity in relation to changes in performance level during a sustained information processing task. *Biological Psychology*, *37*(3), 183–198.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures : Nonreactive research in the social sciences*. Chicago: Rand McNally.

Webb, N. M., & Shavelson, R. J. (1981). Multivariate Generalizability of general educational development ratings. *Journal of Educational Measurement*, *18*(1), 13–22. doi:10.1111/j.1745-3984.1981.tb00839.x

Weber, R., Tamborini, R., Westcott-Baker, A., & Kantor, B. (2009). Theorizing Flow and Media Enjoyment as Cognitive Synchronization of Attentional and Reward Networks. *Communication Theory*, *19*(4), 397–422. doi:10.1111/j.1468-2885.2009.01352.x

Webster, J., & Ahuja, J. S. (2006). Enhancing the design of web navigation systems: The influence of user disorientation on engagement and performance. *MIS Quarterly*, *30*(3), 7. Retrieved from http://aisel.aisnet.org/misq/vol30/iss3/7/

Webster, J., Trevino, L. K., & Ryan, L. (1993). The Dimensionality and Correlates of Flow in Human-Computer Interactions. *Computers in Human Behavior*, *9*, 411–426.

Weissman, L. (1974). *A methodology for studying the psychological complexity of computer programs*. Université de Toronto.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*(1), 49–63.

Wesley, A., Shastri, D., & Pavlidis, I. (2010). A novel method to monitor driver's distractions. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems - CHI EA '10* (p. 4273). New York, New York, USA: ACM Press. doi:10.1145/1753846.1754138

West, P. M., Huber, J., & Sam Min, K. (2004). Altering Experienced Utility: The Impact of Writing and Self-Referencing on Preferences. *Journal of Consumer Research*, *31*(3), 623–631.
Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The Cognitive Walkthrough Method: A Practictioner's Guide. In J. Nielsen & R. Mack (Eds.), *Usability Inspection Methods* (pp. 105–140). New York: John Wiley & Sons.
Whicker, L. M., & Andrews, K. M. (2004). HRM in the Knowledge Economy: Realising the Potential. *Asia Pacific Journal of Human Resources*, *42*(2), 156–165.
White, G. L., & Maltzman, I. (1978). Pupillary activity while listening to verbal passages. *Journal of Research in Personality*, *12*(3), 361–369. doi:10.1016/0092-6566(78)90062-4
White, G. R., Mirza-babaei, P., McAllister, G., & Good, J. (2011). Weak inter-rater reliability in heuristic evaluation of video games. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11* (p. 1441). New York, New York, USA: ACM Press. doi:10.1145/1979742.1979788
White, H. (2000). A reality check for data snooping. *Econometrica*, *68*(5), 1097–1126.
Whittaker, S., Hircshberg, J., Amento, B., Stark, L., Bacchiani, M., Isenhour, P., … Rosenberg, A. (2002). SCANMail: a voicemail interface that makes speech browsable, readable and searchable. In *Proceedings of ACM Conference on Human Factors in Computing 2002* (pp. 275–282). New York, NY: ACM Press.
Wicks, P., Vaughan, T. E., Massagli, M. P., & Heywood, J. (2011). Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nature Biotechnology*, *29*(5), 411–414. doi:10.1038/nbt.1837
Wientjes, C. J. (1992). Respiration in psychophysiology : Methods and applications. *Biological Psychology*, *34*(2-3), 179–203.
Wilkes, Y. (1996). Natural language processing (Introduction to special issue). *Communications of ACM*, *39*(1), 60–62.
Wilson, C. (2006). Triangulation: the explicit use of multiple methods, measures, and approaches for determining core issues in product development. *Interactions*, *13*(6), 46–48. Retrieved from http://portal.acm.org/citation.cfm?id=1167948.1167980
Wilson, M. L., Schraefel, M. c., & White, R. W. (2009). Evaluating advanced search interfaces using established information-seeking models. *Journal of the American Society for Information Science and Technology*, *60*(7), 1407–1422. doi:10.1002/asi.21080
Witvliet, C. V. O., & Vrana, S. R. (1995). Psychophysiological responses as indices of affective dimensions. *Psychophysiology*, *32*, 436–443.
Wixon, D., Holtzblatt, K., & Knox, S. (1990). Contextual design: an emergent view of system design. In *Proceedings of the SIGCHI conference on Human factors in computing systems Empowering people - CHI '90* (pp. 329–336). New York, New York, USA: ACM Press. doi:10.1145/97243.97304
Wood, R. T. a, Griffiths, M. D., & Parke, A. (2007). Experiences of time loss among videogame players: an empirical study. *Cyberpsychology & Behavior : The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, *10*(1), 38–44. doi:10.1089/cpb.2006.9994
Woodman, R. W., Sawyer, J. E., & Griffin, R. W. (1993). Toward a Theory of Organizational Creativity. *The Academy of Management Review*, *18*(2), 293–321.
Woodruff, C. C., Daut, R., Brower, M., & Bragg, A. (2011). Electroencephalographic α-band and β-band correlates of perspective-taking and personal distress. *Cognitive Neuroscience and Neuropsychology*, *22*(15), 744–748. doi:10.1097/WNR.0b013e32834ab439
Wurman, R. S. (1996). *Information architects*. New York: Graphis.
Xu, D. Y., Read, J. C., Sim, G., McManus, B., & Qualter, P. (2009). Children and "smart" technologies : can children's experiences be interpreted and coded ? In *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology (BCS-HCI '09)* (pp. 224–231). Swinton, UK: British Computer Society.
Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012–5015.
Yang, M.-H., & Ahuja, N. (2001). Face Detection and Gesture Recognition for Human–Computer Interaction. In *Kluwer International Series in Video Computing, vol. 1.* Boston: Kluwer.
Yeasin, M., Bullot, B., & Sharma, R. (2006). Recognition of facial expressions and measurement of levels of interest from video. *IEEE Transactions on Multimedia*, *8*(3), 500–508. doi:10.1109/TMM.2006.870737
Yeh, Y. Y., & Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors*, *30*(1), 111–120.
Yockey, H. P. (1992). *Information Theory and Molecular Biology.* Cambridge, U.K: Cambridge University Press.
Young, I., & Veen, J. (2008). *Mental models : aligning design strategy with human behavior.* Brooklyn, N.Y.: Rosenfeld Media.

Yourdon, E. (1978). *Structured walkthroughs*. New York, NY: Yourdon Press.
Youtube. (2012). *Holy Nyans! 60 hours per minute and 4 billion views a day on YouTube*. Retrieved from http://youtube-global.blogspot.fr/2012/01/holy-nyans-60-hours-per-minute-and-4.html
Yu, L., Kittur, A., & Kraut, R. E. (2014a). Distributed analogical idea generation: inventing with crowds. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14* (pp. 1245–1254). doi:10.1145/2556288.2557371
Yu, L., Kittur, A., & Kraut, R. E. (2014b). Searching for analogical ideas with crowds. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14* (pp. 1225–1234). doi:10.1145/2556288.2557378
Yun, C., Shastri, D., Pavlidis, I., & Deng, Z. (2009). O' game, can you feel my frustration?: Improving User's Gaming Experience via StressCam. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI '09* (p. 2195). New York, New York, USA: ACM Press. doi:10.1145/1518701.1519036
Zaman, B., & Abeele, V. V. (2007). How to measure the likeability of tangible interaction with preschoolers. In *Proceedings of CHI Nederland, Vol. 5* (pp. 57–59). Infotec Nederland BV Woerden.
Zaman, B., & Shrimpton-smith, T. (2006). The FaceReader: Measuring instant fun of use. In *Proceeding of NordiCHI 2006* (pp. 457–460).
Zang, N., & Rosson, M. B. (2008). What's in a mashup? And why? Studying the perceptions of web-active end users. In *2008 IEEE Symposium on Visual Languages and Human-Centric Computing* (pp. 31–38). Washington, DC, USA: IEEE. doi:10.1109/VLHCC.2008.4639055
Zeiler, W., Savanovic, P., & Quanjel, E. (2007). Design decision support for the conceptual phase of the design process. In *Conference by the International Association of Societies of Design Research (IASDR'07)*. Hong Kong.
Zhai, S. (2004). Characterizing computer input with Fitts' law para- meters—the information and non-information aspects of pointing. *International Journal of Human–Computer Studies*, *61*, 791–809.
Zhang, L., & Yap, B. (2012). Affect Detection from Text-Based Virtual Improvisation and Emotional Gesture Recognition. *Advances in Human-Computer Interaction*, *2012*, 1–12. doi:10.1155/2012/461247
Zhang, P., Li, N., & Sun, H. (2006). Affective Quality and Cognitive Absorption: Extending Technology Acceptance Research. In *Proceedings of the Hawaii International Conference on System Sciences*. Retrieved from http://www.computer.org/portal/web/csdl/doi/10.1109/HICSS.2006.39
Zhang, Y., Bulling, A., & Gellersen, H. (2012). Towards pervasive eye tracking using low-level image features. In *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12* (p. 261). New York, New York, USA: ACM Press. doi:10.1145/2168556.2168611
Zhang, Z., Basili, V., & Shneiderman, B. (1999). Perspective-based usability inspection: An empirical validation of efficacy. *Empirical Software Engineering*, *4*(1), 43–69. doi:10.1023/A:1009803214692
Zhao, Z., Badam, S. K., Chandrasegaran, S., Park, D. G., Elmqvist, N. L. E., Kisselburgh, L., & Ramani, K. (2014). skWiki: a multimedia sketching system for collaborative creativity. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14* (pp. 1235–1244). doi:10.1145/2556288.2557394
Zuboff, S. (1988). *Smart machine: The future of work and power*. New York: Basic Books.
Zuboff, S. (2010). Creating Value in the Age of Distributed Capitalism. *McKinsey Quarterly*, 45–55.

Table des figures et tableaux