Développement et évaluation de méthodes permettant d'étudier les sources d'hétérogénéité du DIF et du Response Shift lors de l'analyse des données rapportées par les patients

Yseulys Dubuy

DOCTORAT BIOLOGIE BRETAGNE SANTE LOIRE

Nantes Université

THESE DE DOCTORAT DE

NANTES UNIVERSITE

ECOLE DOCTORALE N° 605
Biologie Santé
Spécialité : *Epidémiologie, Analyse de Risque, Recherche Clinique*

Par
Yseulys DUBUY

Développement et évaluation de méthodes permettant d'étudier les sources d'hétérogénéité du DIF et du Response Shift lors de l'analyse des données rapportées par les patients

Thèse présentée et soutenue à Nantes, le 29 novembre 2022
Unité de recherche : UMR INSERM 1246 SPHERE « methodS in Patient-centered outcomes and HEalth ResEarch »

Rapporteurs avant soutenance :

Nancy E. MAYO	Professeur, Université McGill
Alexandra ROUQUETTE	PU-PH, Université Paris-Saclay

Composition du Jury :

Président :	Alain LEPLEGE	PU, Université Paris Cité
Directeur de thèse :	Jean-Benoit HARDOUIN	MCU-PH, Nantes Université
Co-directrice de thèse :	Véronique SEBILLE	PU-PH, Nantes Université
Co-encadrante de thèse :	Myriam BLANCHIN	IR, Nantes Université

Invité
Jean-Luc KOP MCU, Université de Lorraine

After the flood all the colors came out...

It was a beautiful day, don't let it get away

U2

Remerciements

Et voilà... Il est maintenant temps de remercier toutes les personnes qui se sont retrouvées embarquées dans mon aventure de doctorante, de façon volontaire ou involontaire, et pour le meilleur comme le pire !

Pour commencer ces remerciements, tout d'abord un immense merci à mes super directeurs de thèse Jean-Benoit, Véronique et Myriam.

Jean-Benoit, tout d'abord merci d'avoir répondu en 2017 à mes mails de petite étudiante à la recherche d'un Master de biostatistique, de m'avoir proposé un stage en Master 2 avec vous trois, puis d'avoir monté des dossiers pour chercher un financement de thèse. Un très grand merci également pour tous tes bons conseils au cours de cette thèse, ta gentillesse et bien sûr pour toujours nous faire rigoler en réunion.

Véronique, merci "++++", pour tout : ta bienveillance, ton soutien social (je me rappellerai longtemps de cette quasi-perte d'ESM sur mon PC...), ta confiance en moi, nos discussions pédagogiques, scientifiques... Bref, MERCI pour tout ce que tu fais pour moi !

Enfin, Myriam, un grand grand merci pour ta gentillesse, ton écoute, ton implication sans faille dans cette thèse, pour m'impliquer dans tes projets (Team DataViz), pour ton œil de Lynx en relecture, et puis bien sûr, pour toujours m'écouter alors que plus personne n'est concentré en réunion le vendredi après-midi...;-).

Bref, merci à tous les trois pour tout, car sans vous, rien n'aurait été possible.

Merci aussi aux membres de mon jury de thèse d'avoir accepté de juger mon travail : Alexandra Rouquette, Nancy Mayo, Alain Leplège et Jean-Luc Kop. Merci pour votre bienveillance et pour tous nos échanges au cours de ma soutenance. Un merci tout particulier à Jean-Luc qui était également membre de mon comité de suivi individuel. J'en profite également pour remercier les deux autres membres de mon CSI, Antoine Vannier et Anne Congard.

J'en viens à remercier mes super collègues pour leur soutien : Marie (ma chère co-bureau), Bastien (qui a le droit à toutes mes « ronchonnades » et mes mauvaises blagues), Elodie (merci infiniment pour tes relectures de toute dernière minute), Angely (Team partenaires de galère), Odile, Marianne (la doctorante), ... et tous les autres ! Merci pour vos encouragements, notamment ces dernières semaines. Merci aussi à ceux qui sont partis pour de nouvelles aventures, mais qui ont toujours un gentil mot en stock : Line, Jeanne, Arthur et Camille. Pour clôturer le versant "pro" de ces remerciements, merci à Marianne (la psychologue), pour tous nos échanges, ton soutien et pour ta disponibilité. Ton aide sur le PTG m'a été très précieuse.

Pour terminer, un merci (infini) à mes proches, ma famille de sang et de cœur. Merci à ma petite Sharon, sans qui mon arrivée à Nantes n'aurait pas été la même ! À ma cousine Audrey, pour tous ses précieux conseils sur la thèse, et pour l'hébergement parisien. Merci à mon parrain, que je n'ai pas vu assez avec cette thèse, mais ce n'est que partie remise. Enfin, merci à mon compagnon Thibaud, qui a connu tous les hauts (et les bas !) de ces trois années de doctorat. Merci pour ton soutien, pour ta patience, et pour croire en moi comme tu le fais.

Cette thèse est pour mes parents :

Maman et Eric, merci pour votre soutien à toute épreuve. Merci de m'avoir toujours encouragée, de m'avoir toujours poussée à faire mieux, d'avoir été là à chaque fois que j'en avais besoin. Sans vous, je n'aurais jamais réussi à arriver jusqu'ici ♡.

Papa, cette année en décembre, ça fera dix ans que tu es parti... Où que tu sois, j'espère que tu es fier de moi.

Valorisations scientifiques

Les auteurs ayant une contribution équivalente à une publication sont indiqués par un astérisque.

Publications issues de la thèse

- **Dubuy Y**, Sébille V, Grall-Bronnec M, Challet-Bouju G, Blanchin M, Hardouin JB. Evaluation of the link between the Guttman errors and response shift at the individual level (2021). *Quality of Life Reseach.* https://doi.org/10.1007/s11136-021-03015-9
- **Dubuy Y**, Sébille V, Bourdon M, Hardouin JB, Blanchin M. Posttraumatic growth inventory : challenges with its validation among French cancer patients (2022). *BMC Medical Research Methodology.* https://doi.org/10.1186/s12874-022-01722-6
- **Dubuy Y**, Hardouin JB, Blanchin M.*, Sébille V.* Identification of sources of DIF using covariates : a simulation study comparing two approaches based on Rasch family models. *[Soumis]*

Autres publications en lien avec la thèse

- Sebille V*, **Dubuy Y***, Feuillet F, Cinotti R, Roquilly A. Does Differential Item Functioning jeopardizes the comparability of health-related quality of life assessment between patients and proxies in moderate to severe traumatic brain injury patients ? *[Soumis]*

Communications orales

Dans un colloque international

- **Dubuy Y**, Sebille V, Blanchin M, Hardouin JB (2020). Assessment of the link between response shift and the change over time in the number of Guttman errors via a simulation study. *27th Annual Conference of the International Society for Quality of Life Research (ISOQOL)*, Virtual Event, 19 – 23 octobre 2020.

Dans un colloque national

- **Dubuy Y**, Sebille V, Blanchin M, Hardouin JB (2020). Détection du *response shift* au niveau individuel à l'aide des erreurs de Guttman. *14ème Conférence Francophone d'Epidémiologie Clinique (EPICLIN)*, e-conférence, 15 – 16 septembre 2020.

Communications affichées

Dans un colloque international

- **Dubuy Y**, Sebille V, Blanchin M, Hardouin JB (2019). Validation of a Guttman errors-based method to detect response shift at an individual level. *26th Annual Conference of the International Society for Quality of Life Research (ISOQOL)*, San Diego, California, United States, 20 – 23 october 2019.
- **Dubuy Y**, Sebille V, Hardouin JB, Blanchin M (2021). Assessment of the heterogeneity of DIF using covariates with Rasch family models. *28th Annual Conference of the International Society for Quality of Life Research (ISOQOL)*, Virtual Event, 12 – 28 octobre 2021.
- **Dubuy Y**, Sebille V, Bourdon M, Blanchin M (2021). Validation of the French version of the Post-traumatic Growth Inventory in melanoma and breast cancer patients. *28th Annual Conference of the International Society for Quality of Life Research (ISOQOL)*, Virtual Event, 12 – 28 octobre 2021.

Dans un colloque national

- Dubuy Y, Sebille V, Blanchin M, Hardouin JB (2019). Validation d'une méthode de détection du response shift, au niveau individuel à l'aide des erreurs de Guttman. Journées 2019 du GDR « Statistiques & Santé », de la Société Française de Biométrie et du groupe « Biopharmacie » de la Société Française de Statistique, Paris, France, 10 – 11 octobre 2019.
- Dubuy Y, Sebille V, Blanchin M, Hardouin JB (2020). Version française du « Post-traumatic growth inventory » : écueils de traduction et impact sur les propriétés psychométriques *16ème Conférence Francophone d'Epidémiologie Clinique (EPICLIN)*, e-conférence, 15 – 16 septembre 2020.

Table des matières

Table des figures

Liste des tableaux

Abréviations

2-PLM	Modèle logistique à deux paramètres (*Two-Parameter Logistic Model*)
3-PLM	Modèle logistique à trois paramètres (*Three-Parameter Logistic Model*)
AIC	Critère d'information d'Akaike (*Akaike Information Criterion*)
AIDS	Syndrome d'immunodéficience acquise (*Acquired ImmunoDeficiency Syndrome*)
AL	Appréciation de la vie (*Appreciation of Life*)
AUROC	Aire sous la courbe ROC (*Area Under the ROC curve*)
BIC	Critère d'information bayésien (*Bayesian Information Criterion*)
BMI	Indice de masse corporelle (*Body Mass Index*)
BVR	Résidus bivariés (*BiVariate Residuals*)
CAIC	Critère d'information d'Akaike consistant (*Consistent Akaike Information Criterion*)
CART	*Classification And Regression Trees*
CCC	Courbe caractéristique des modalités de réponse (*Category Characteristic Curves*)
CFA	Analyse factorielle confirmatoire (*Confirmatory Factor Analysis*)
CFI	*Comparative Fit Index*
CLV	*Clustering of variables around Latent Variables*
COSMIN	*COnsensus-based Standards for the selection of health Measurement INstruments*
CTT	Théorie classique des tests (*Classical Test Theory*)
DIF	Fonctionnement différentiel des items (*Differential Item Functioning*)
DIF H	DIF Homogène
DIF NH	DIF Non homogène
DSF	*Differential Step Functioning*
DSM	Manuel diagnostique et statistique des troubles mentaux, et des troubles psychiatriques (*Diagnostic and Statistical Manual of Mental Disorders*)
ED	Troubles du comportement alimentaire (*Eating Disorders*)
EDI-2	Inventaire des troubles du comportement alimentaire version 2 (*Eating Disorder Inventory version 2*)
EG	Erreurs de Guttman
FPR	Taux de faux positifs (*False Positive Rate*)
GE	Erreurs de Guttman (*Guttman Errors*)
GMM	Modèles mixtes à classes latentes (*Growth Mixture Model*)
GPCM	Modèle de crédit partiel généralisé (*Generalized Partial Credit Model*)

HIV	Virus de l'immunodéficience humaine (*Human Immunodeficiency Virus*)
HRQoL	Qualité de vie liée à la santé (*Health-Related Quality of Life*)
ICC	Courbe caractéristique des items (*Item Characteristic Curve*)
IFT	*Item-Focused Tree*
INFIT	*Inlier-sensitive fit*
IRF	Fonction de réponse à l'item (*Item Response Function*)
IRT	Théorie de la réponse aux items (*Item Response Theory*)
IRT-C	*Item Reponse Theory with Covariate*
LISREL	*LInear Structural RELationships*
LPCM	Modèle de crédit partiel longitudinal (*Longitudinal Partial Credit Model*)
LRT	Test du rapport de vraisemblance (*Likelihood-Ratio Test*)
LVMM	Modèle à variable latente de mélange (*Latent Variable Mixture Model*)
MD	Données manquantes (*Missing Data*)
MIMIC	*Multiple Indicators Multiple Causes*
MV	Maximum de vraisemblance
NP	Nouvelles opportunités (*New Possibilities*)
NUR	Recalibration non uniforme (*Non-Uniform Recalibration*)
OUTFIT	*Outlier-sensitive fit*
PCA	Analyse en composantes principales (*Principal Component Analysis*)
PCM	Modèle de crédit partiel (*Partial Credit Model*)
PGI	*Patient Generated Index*
PRO	Résultat rapporté par les patients (*Patient-reported outcome*)
PS	Force personnelle (*Personal Strength*)
PTG	Développement post-traumatique (*Posttraumatic Growth*)
PTGI	Inventaire du développement post-traumatique (*Posttraumatic Growth Inventory*)
QoLAP	*Quality of Life Appraisal Profile*
RC	Recalibration
RC NU	Recalibration non uniforme
RC U	Recalibration uniforme
RMSEA	*Root-Mean-Square Error of Approximation*
RMT	Théorie de la mesure de Rasch (*Rasch Measurement Theory*)
RO	Relations aux autres (*Relating to Others*)
ROC	*Receiver Operating Characteristic*
ROSALI	*RespOnse Shift ALgorithm at Item-level*
ROSALI-IRT	*RespOnse Shift ALgorithm at the Item level based on Item response Theory*

ROSALI-RMT	*RespOnse Shift ALgorithm at the Item level based on Rasch Measurement Theory*
RS	*Response shift*
SC	Changement spirituel (*Spiritual Change*)
SEIQOL	*Schedule for the Evaluation of Individual Quality of Life*
SEM	Modèles à équations structurelles (*Structural Equation Modeling*)
SRMR	*Standardized Root-Mean-square Residual*
TPR	Taux de vrais positifs (*True Positive Rate*)
UR	Recalibration uniforme (*Uniform Recalibration*)

Chapitre 1

Introduction

Les *Patient-Reported Outcomes* (PRO)

La mesure de concepts subjectifs - comme la qualité de vie, la fatigue ou encore les symptômes dépressifs - est capitale pour évaluer l'état de santé des patients, en prenant en compte leurs perspectives. En effet, cela fait maintenant longtemps que la définition de la santé n'est plus seulement réduite à la seule absence de maladie ou d'infirmité, mais prend également en compte des aspects comme le bien-être physique, mental et social [1].

Ces aspects de l'état de santé des patients, peuvent être obtenus en "écoutant la voix" des patients. On parle alors de résultats rapportés par les patients (*Patient-reported outcomes*, PRO). Plus précisément, les PRO sont définis comme "toute mesure de l'état de santé d'un patient rapportée directement par lui-même, sans interprétation d'un médecin ou de toute autre personne". Ils permettent de s'intéresser à la façon dont un patient se sent (ou fonctionne) par rapport à son état de santé [2]. Les PRO permettent d'obtenir, grâce aux patients, des informations uniques sur l'impact d'une maladie ou d'un traitement, qui ne pourraient pas être obtenues autrement qu'en prenant en compte leur point de vue [3]. En pratique, le terme PRO est un terme générique qui peut par exemple faire référence [4] :

- Aux symptômes d'une maladie ou aux effets secondaires d'un traitement, comme la douleur, la fatigue physique, la fatigue mentale, ou encore l'anxiété ;
- Aux fonctionnements des individus, tels que le fonctionnement physique, sexuel, social, émotionnel ou cognitif ;
- À des constructions multidimensionnelles telles que la qualité de vie liée à la santé (*Health-Related Quality of Life*, HRQoL).

Les PRO sont de plus en plus utilisés dans les essais cliniques [4] et ils figurent aujourd'hui parmi les critères principaux de certaines études (notamment en oncologie [5], dans le champ de la santé mentale [6], en soins de réadaptation et en soins palliatifs [4]). L'un des PRO les plus utilisés en recherche clinique est la qualité de vie liée à la santé, qui permet de s'intéresser à l'impact d'une maladie ou d'une intervention sur la qualité de vie des patients (tel que perçu par ces derniers). La notion de PRO est néanmoins plus large, puisqu'elle englobe tous les aspects de la santé qui peuvent être rapportés par le patient lui-même.

Patient-Reported Outcomes : pourquoi sont-ils si importants ?

Afin d'étudier les effets d'une intervention de santé, la recherche biomédicale s'est longtemps appuyée sur des critères cliniques (comme la survie ou la progression d'une maladie). Néanmoins, avec les progrès de la médecine, l'objectif thérapeutique est repoussé toujours plus loin, et on vise alors parfois à améliorer des critères physiologiques très spécifiques, qui ne se traduisent pas nécessairement par des résultats bénéfiques tangibles pour le patient [2]. Les PRO sont ainsi apparus dans les essais cliniques de phase III (visant à étudier l'efficacité de nouveau traitement ou d'une nouvelle intervention) suite aux recommandations de l'administration américaine FDA (*Food and Drug Administration*), parues en 2006 [2].

L'intérêt autour des PRO s'explique par le fait qu'ils sont centrés sur les points de vue et les expériences des patients, et permettent ainsi d'obtenir une vision globale de l'état de santé de l'individu, qui va au-delà des critères cliniques comme la survie, la progression de la maladie ou le niveau d'un biomarqueur [7, 8]. Leur utilisation croissante est en fait multifactorielle :

– On peut tout d'abord évoquer les progrès de la médecine et le vieillissement de la population (dans les pays occidentaux), qui ont entraîné une augmentation du nombre d'individus vivant avec une maladie chronique [9, 10]. Par exemple, en 2015, l'Assurance Maladie a estimé que 35% de la population française pourrait être concernée par ce type de pathologie [11]. Pour les patients atteints d'une maladie chronique, les PRO peuvent être utilisés comme des indicateurs de suivi de leur état de santé. Ils peuvent également permettre d'intégrer les attentes des patients dans la gestion thérapeutique de leur maladie (pour que les objectifs thérapeutiques soient en phase avec les objectifs des patients).

– De plus, pour certaines pathologies comme le cancer, il est de plus en plus difficile de trouver de nouveaux traitements entrainant une amélioration substantielle des critères dits objectifs comme la survie [12]. L'utilisation des PRO peut dans ce cas permettre d'arbitrer des thérapies présentant des résultats similaires en termes de survie, en sélectionnant celle assurant la meilleure qualité de vie ou le meilleur bien-être aux patients.

– Par ailleurs, de nombreux symptômes et niveaux de fonctionnement ne sont pas directement mesurables. Les patients représentent alors la seule "source d'information". C'est notamment le cas lorsque l'on s'intéresse à la douleur, la fatigue, le fonctionnement émotionnel, etc. [2, 7, 13].

– Enfin, il y a aujourd'hui une véritable transformation du rôle des patients, en recherche et en pratique clinique, avec la volonté de prendre en compte leurs préférences et les impliquer dans la prise de décision. On cherche aujourd'hui à s'intéresser à ce qui compte vraiment pour eux, notamment grâce au développement des *Core Outcomes Sets*, parmi lesquels figurent de nombreux PRO [14].

Patient-Reported Outcomes **: Recueil des informations auprès des patients**

Différentes approches peuvent être utilisées pour recueillir le point de vue des patients, parmi lesquelles les questionnaires et les entretiens (à condition que la personne menant l'entretien ne soit là que pour recueillir les réponses des patients et qu'il/elle n'interfère pas avec ces dernières) [12]. Ces informations sont le plus souvent recueillies par l'intermédiaire de questionnaires standardisés et ayant été validés pour s'assurer qu'ils mesurent bien le concept (ou le "construit" en termes psychométriques) qu'ils sont censés mesurer de manière fiable. Ces questionnaires peuvent comporter un ou plusieurs items (*i.e.,* des questions), qui peuvent être regroupés en une ou plusieurs dimensions. Lorsque le questionnaire est composé de plusieurs dimensions, on dit qu'il est multidimensionnel. Chaque dimension représente alors un concept parmi un concept multidimensionnel. Par exemple, le fonctionnement physique est un concept inclus dans le concept multidimensionnel qu'est la qualité de vie liée à la santé. Les formats de réponse possibles pour les items sont multiples et peuvent varier d'un item à l'autre au sein d'un même questionnaire. Les données issues de ces questionnaires sont désignées par l'expression : "données rapportées par les patients".

L'analyse des données rapportées par les patients

Les données rapportées par les patients diffèrent des données dites "objectives" puisque la mesure que l'on obtient pour chaque individu peut être influencée par les opinions et les références de la personne qui les rapporte (même si ce n'est pas systématique). Si c'est le cas, on parle alors de données subjectives.

L'analyse moderne des données rapportées par les patients suppose que les réponses aux items d'un questionnaire sont le reflet des construits non observables ciblés par ce questionnaire. On parle de paradigme réflectif. Ces réponses sont utilisées pour :

(i) quantifier le niveau des individus pour le construit étudié ;

(ii) ordonner les individus les uns par rapport aux autres en fonction de leur niveau.

Les quatre principales méthodes d'analyse pour ce type de données sont : la théorie classique des tests (*Classical Test Theory*, CTT), les modèles à équations structurelles (*Structural Equation Modeling*, SEM), les modèles de la famille de Rasch (*Rasch Measurement Theory*, RMT) et les modèles de la famille de Lord, aussi appelés modèles de la théorie de la réponse aux items (*Item Response Theory*, IRT). Lors de l'analyse de ce type de données, de nombreuses difficultés peuvent émerger. Parmi elles, on peut notamment citer la violation de l'hypothèse de l'invariance de la mesure [15].

L'hypothèse de l'invariance de la mesure

Les méthodes d'analyse qui viennent d'être mentionnées reposent sur l'hypothèse de l'invariance de la mesure [15]. Pour que cette hypothèse soit vérifiée, il faut que :

- Les individus ayant des niveaux de construit très proches répondent de façon similaire aux items du questionnaire, et ce, indépendamment de leurs caractéristiques. On parle d'invariance entre groupes d'individus.
- Les réponses d'un même individu entre les deux temps de mesure reflètent bien l'évolution du construit ciblé : c'est-à-dire des réponses similaires si le niveau du construit est stable et des réponses qui s'améliorent (ou se détériorent) si le niveau du construit s'améliore (ou se détériore) au cours du temps. On parle alors d'invariance longitudinale.

En pratique, il peut néanmoins y avoir des différences dans la façon dont les individus interprètent le construit ciblé par le questionnaire et les items associés, du fait de leurs caractéristiques culturelles, environnementales, personnelles, etc. [16]. Ces différences sont connues sous le nom de fonctionnement différentiel des items (*Differential Item Functioning*, DIF). De façon similaire, un individu confronté à un événement de santé majeur, ayant des répercussions importantes sur sa vie, peut être amené à ne plus interpréter de la même façon les questions qui lui sont posées. Sa vision du construit ciblé peut être modifiée, l'importance qu'il accorde aux différentes dimensions de ce construit peut changer, et ses normes de mesures internes peuvent évoluer [17]. Ces changements longitudinaux peuvent entraîner une dissonance entre l'évolution rapportée par le patient et l'évolution du construit ciblé. Ce phénomène est connu sous le nom de *response shift* (RS) et pourrait être associé au fait que les patients s'adaptent ou non aux événements auxquels ils font face [18].

Le DIF et le *response shift* doivent être pris en compte lors de l'analyse des données autorapportées par les patients puisqu'ils peuvent entraîner une mésinterprétation des mesures obtenues, pouvant conduire à des conclusions potentiellement erronées [16]. Par ailleurs, ils représentent également des critères d'intérêt en eux-mêmes, pour une interprétation plus fine des différences entre les individus et des changements expérimentés. Plusieurs méthodes existent pour détecter et prendre en compte ces phénomènes. Il y a néanmoins aujourd'hui une véritable volonté d'aller plus loin et tenter non pas simplement de les détecter, mais également les expliquer :

- Pour le DIF, l'une des thématiques de recherche actuelle est la prise en compte simultanée de plusieurs covariables afin de tenter de démêler leurs effets sur le fonctionnement différentiel des items [19–22].
- Pour le *response shift*, l'une des pistes de recherche actuelle est l'étude de sa variabilité interindividuelle (c'est-à-dire son hétérogénéité) [23]. En effet, face à un même événement, les individus n'expérimentent pas tous du *response shift*. Par ailleurs, pour ceux qui en expérimentent, ce n'est pas forcément de la même façon. Explorer l'hétérogénéité du*response shift* pourrait permettre de mieux comprendre le phénomène et, dans un champ plus large, mieux comprendre l'adaptation des individus face aux événements de santé qu'ils expérimentent.

À terme, cela pourrait permettre d'identifier des leviers d'action pour accompagner au mieux les patients [23].

Objectifs et plan du manuscrit

Les travaux présentés dans ce manuscrit s'inscrivent dans la lignée de ces axes de recherche. Ils ont été initialement entrepris pour tenter de prendre en compte la variabilité interindividuelle du *response shift* à l'aide d'indicateurs calculés directement à partir des réponses des individus. Ce type de méthode pourrait permettre d'identifier les caractéristiques des patients associées au *response shift* sans hypothèses faites *a priori*. Une autre piste permettant d'étudier la variabilité interindividuelle du *response shift* (et tenter d'expliquer ce phénomène) est l'intégration de covariables lors de la détection du *response shift*. Pour ce faire, ces covariables doivent être identifiées sur la base d'hypothèses cliniques formulées *a priori*. Néanmoins, pour pouvoir étudier les changements dans l'interprétation des items d'un questionnaire au cours du temps en considérant plusieurs covariables, il faut tout d'abord être capable de prendre correctement en compte les différences d'interprétation qui pourraient être induites par ces covariables au premier temps de mesure. C'est pourquoi une partie de mes travaux de thèse se sont focalisés sur l'étude du DIF en présence de plusieurs covariables. Enfin, le *response shift* est présenté dans la littérature comme pouvant être associé au développement post-traumatique (correspondant au fait qu'un individu puisse percevoir des changements psychologiques positifs suite à un événement défiant hautement ses ressources [24]). Plus précisément, le développement post-traumatique pourrait être l'une des causes de la survenue de *response shift* [18]. Aussi, il pourrait être intéressant de s'intéresser aux liens entre ces deux phénomènes et déterminer, par exemple, si l'on retrouve du *response shift* chez des individus ayant expérimenté du développement post-traumatique. Pour pouvoir mener à bien ce type de recherche, il est nécessaire d'avoir un outil de mesure fiable et valide du développement post-traumatique. L'outil de mesure le plus connu est l'inventaire du développement post-traumatique (développé en anglais américain par les chercheurs à l'origine du concept) [24]. En France, si cet inventaire a été traduit il y a de ça plus de dix ans, les données permettant d'assurer sa validité de construit sont manquantes. C'est pourquoi le dernier travail de cette thèse porte sur l'évaluation des propriétés psychométriques de ce questionnaire.

Ce manuscrit est organisé en deux grandes parties :

– La première partie comporte deux chapitres qui présentent les différentes notions et méthodes utiles pour la lecture des chapitres suivants. Concrètement, le premier chapitre présente les différents modèles à variable latente utiles à l'analyse des données rapportées par les patients et le second chapitre est consacré aux notions de DIF et de *response shift*.

– La seconde partie est composée de trois chapitres détaillant mes travaux de thèse. Le premier chapitre présente les travaux réalisés dans l'objectif d'individualiser la détection du *response shift* entre deux temps de mesure à l'aide des erreurs de Guttman. Le deuxième chapitre porte sur les développements méthodologiques proposés pour étudier le DIF en présence de deux covariables binaires et les simulations réalisées pour les évaluer. Enfin, le troisième chapitre présente l'étude des propriétés psychométriques d'une des versions françaises de l'inventaire du développement post-traumatique. Chacun de ces travaux fait l'objet d'une discussion présentée dans le chapitre concerné. Cette seconde partie s'achève par une discussion plus générale, synthétisant nos principaux constats et décrivant les perspectives en lien avec ces travaux.

Partie I : État des connaissances

Chapitre 2

Modèles à variables latentes pour l'analyse des données auto-rapportées

Sommaire

2.1 Qu'est-ce qu'une variable latente ? (paradigme réflectif)

Cette section est basée sur la définition issue du dictionnaire de l'Association américaine de psychologie (APA) [25].

En psychométrie, le terme "variable latente" désigne un construit théorique qui est utilisé pour expliquer une ou plusieurs variables observées (paradigme réflectif). Ces variables sont dites "latentes" car elles ne peuvent pas être directement observées ou mesurées (comme ça pourrait être le cas de l'âge ou de la température corporelle). À la place, on les approxime par l'intermédiaire de différentes mesures observées qui sont supposées les évaluer. Par exemple, considérons un chercheur souhaitant étudier la dépression. Comme la dépression est un construit qui ne peut pas être directement évalué, ce chercheur va utiliser (ou développer) un questionnaire contenant des items s'intéressant aux manifestations de la dépression (comme un sentiment de tristesse, de découragement, de solitude, ou encore une attitude pessimiste ou un éloignement vis-à-vis des autres). Les réponses des participants peuvent ensuite être analysées pour ordonner les individus les uns par rapport aux autres vis-à-vis du construit d'intérêt (la dépression).

2.2 Modèles à équations structurelles

Les modèles à équations structurelles (*Structural Equation Models*, SEM) sont des outils statistiques flexibles qui permettent de modéliser et de tester les relations hypothétiques qui peuvent s'établir entre des variables latentes (non directement observables) et des variables dites "manifestes" (c.-à-d., que l'on peut observer, comme les réponses aux items d'un questionnaire). Ces modèles ont été popularisés dans les années 1970 par Karl Jöreskog, notamment grâce au modèle LISREL (*LInear Structural RELationships*) et au logiciel éponyme [26]. Toute cette section sur les modèles à équations structurelles se restreint aux SEM réflectifs, où l'on suppose que les variables latentes sont des construits non observables, mais estimables à partir de variables manifestes qui en reflètent les effets (hypothèse souvent faite dans le cadre des données rapportées par les patients).

2.2.1 Spécification des relations entre les variables latentes et manifestes

Lorsque l'on s'intéresse à modéliser des liens entre des variables latentes et des variables manifestes, il est nécessaire de proposer un modèle *a priori*, où l'on définit les relations causales présumées entre les différents types de variables. On parle de spécification du modèle. Cette spécification est censée être guidée par la théorie, la littérature, ou encore des hypothèses que l'on souhaite tester. Dans les SEM, on distingue en fait deux sous-modèles : (i) le **modèle de mesure** qui relie les variables latentes aux variables manifestes et (ii) le **modèle structurel** qui lie les variables latentes entre elles.

2.2.2 Représentation graphique des SEM

Dans le cadre des SEM, les relations présumées entre les variables latentes et les variables observées sont souvent représentées graphiquement. La figure 2.1 est la représentation graphique d'un SEM où :

- Trois variables manifestes (x_{11}, x_{21} et x_{31}) sont supposées représenter la variable latente ξ_1 ;
- Deux variables manifestes (x_{12} et x_{22}) sont utilisées pour représenter la variable latente ξ_2 ;
- Cinq variables manifestes (y_1, y_2, y_3, y_4 et y_5) sont utilisées pour représenter la variable latente η ;
- Les variables latentes ξ_1 et ξ_2 sont expliquées par la variable latente η (les variables ξ_1 et ξ_2 sont dites endogènes) ;
- La variable latente η n'est pas expliquée par d'autres variables au sein du modèle (cette variable est dite exogène).

Dans l'exemple du SEM spécifié dans la figure 2.1, on peut distinguer trois modèles de mesure (mis en évidence par des encadrés verts) : le modèle de mesure pour ξ_1 qui relie la variable latente endogène ξ_1 aux variables manifestes (x_{11}, x_{21}, x_{31}), le modèle de mesure pour ξ_2 reliant la variable latente endogène ξ_2 aux variables manifestes (x_{12}, x_{22}) et le modèle de mesure pour η reliant la variable latente exogène η aux cinq variables manifestes (y_1, y_2, y_3, y_4, y_5). Dans ces modèles de mesure, les liens unissant les variables latentes et les variables manifestes sont appelés "*factor loadings*" et notés λ (suivi de l'indice de la variable manifeste à laquelle ils sont

associés). Les termes désignés par les lettres ε et δ (et indicés en fonction de la variable manifeste à laquelle ils se rapportent) sont appelés "termes d'erreur". Leur variance représente la quantité de variation de la variable manifeste qui est due à une erreur de mesure ou qui n'est pas expliquée par la variable latente. Ces termes sont également parfois appelés "résidus" ou "variable d'erreur".

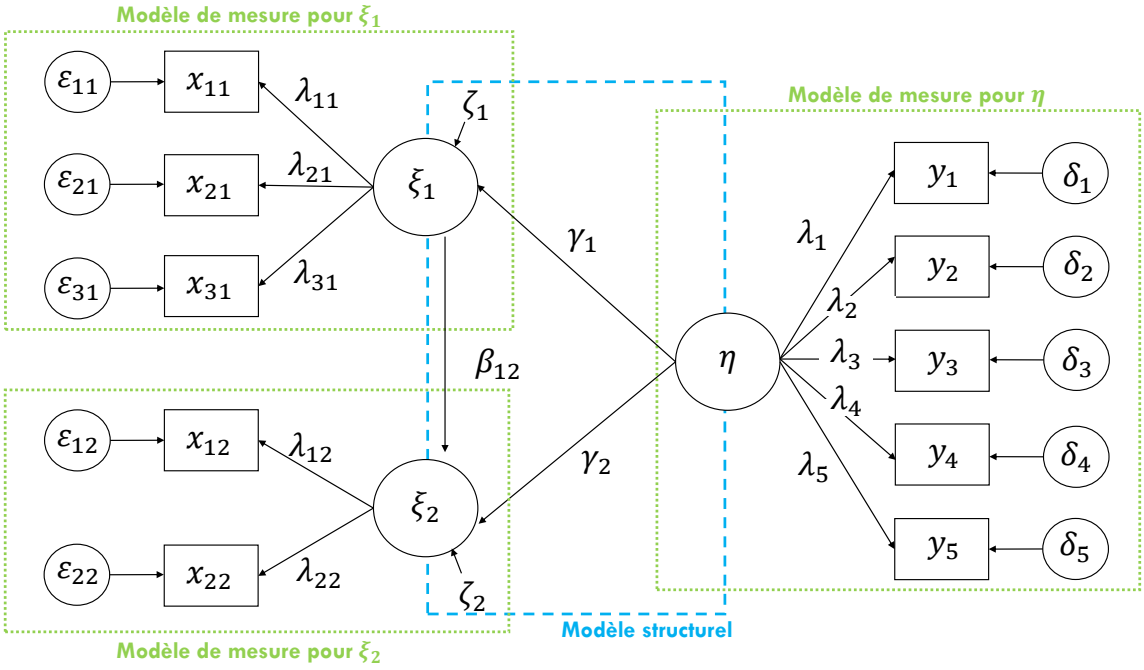


FIGURE 2.1 – Représentation graphique d'un modèle à équations structurelles

Les relations hypothétiques entre les variables latentes forment le modèle structurel du SEM. Dans le SEM représenté dans la figure 2.1, le modèle structurel est encadré en bleu. On y observe les effets de la variable latente exogène η sur les variables latentes endogènes ξ_1 et ξ_2 (représentés par les flèches nommées γ_1 et γ_2) et l'effet de la variable latente endogène ξ_1 sur la variable latente endogène ξ_2 (représenté par la flèche β_{12}). Les paramètres spécifiant les liens entre les variables latentes sont appelés "coefficients structurels". Des résidus latents ζ_1 et ζ_2 sont également introduits dans ce modèle structurel (associés respectivement aux variables latentes ξ_1 et ξ_2). Leur variance représente la quantité de variation des variables latentes ξ_1 et ξ_2 inexpliquée par les variables sur lesquelles ξ_1 et ξ_2 sont régressées. Dans la littérature, ces résidus latents sont désignés sous le nom de "termes de perturbation latente" (*latent disturbance terms*) ou de "perturbations structurelles" (*structural disturbances*)

2.2.3 Formulation mathématique des SEM

Mathématiquement, un SEM correspond à un système d'équations traduisant les relations causales attendues entre les variables latentes et les variables manifestes et entre les différentes variables latentes. Les équations de ce système vont donc dépendre des liens que l'on suppose entre les variables. Par exemple, les modèles de mesure et le modèle structurel du SEM représenté dans la figure 2.1 s'écrivent [1] :

– Modèle de mesure pour ξ_1 :

$$\begin{cases} x_{11_i} = \tau_{11} + \lambda_{11} \times \xi_{1_i} + \varepsilon_{11_i} \\ x_{21_i} = \tau_{21} + \lambda_{21} \times \xi_{1_i} + \varepsilon_{21_i} \\ x_{31_i} = \tau_{31} + \lambda_{31} \times \xi_{1_i} + \varepsilon_{31_i} \end{cases} \tag{2.1}$$

– Modèle de mesure pour ξ_2 :

$$\begin{cases} x_{12_i} = \tau_{12} + \lambda_{12} \times \xi_{2_i} + \varepsilon_{12_i} \\ x_{22_i} = \tau_{22} + \lambda_{22} \times \xi_{2_i} + \varepsilon_{22_i} \end{cases} \tag{2.2}$$

– Modèle de mesure pour η :

$$\begin{cases} y_{1_i} = \tau_1 + \lambda_1 \times \eta_i + \delta_{1_i} \\ \dots \\ y_{5_i} = \tau_5 + \lambda_5 \times \eta_i + \delta_{5_i} \end{cases} \tag{2.3}$$

– Modèle structurel :

$$\begin{cases} \xi_{1_i} = \tau_{\xi_1} + \gamma_1 \times \eta_i + \zeta_{1_i} \\ \xi_{2_i} = \tau_{\xi_2} + \beta_{12} \times \xi_{1_i} + \gamma_2 \times \eta_i + \zeta_{2_i} \end{cases} \tag{2.4}$$

Dans chaque équation, on remarque la présence d'un *intercept* noté τ et indicé en fonction de la variable à laquelle il se rapporte. Ces coefficients sont rarement représentés dans les *path diagram* (les représentations graphiques de SEM). L'*intercept* d'une variable manifeste représente la valeur moyenne attendue pour cette variable si la variable latente est nulle. Il y a également des *intercepts* pour les variables latentes endogènes. L'ensemble des notations considérées dans cet exemple est synthétisé dans le tableau 2.1.

1. Dans ces équations, l'indice i désigne l'individu i.

Notation	Correspondance
η	Variable latente exogène
ξ_1 et ξ_2	Variables latentes endogènes
x_{11}, x_{21}, x_{31}	Variables manifestes liées à la variable latente endogène ξ_1
x_{12}, x_{22}	Variables manifestes liées à la variable latente endogène ξ_2
y_1 à y_5	Variables manifestes liées à la variable latente exogène η
τ_{11}, τ_{21}, τ_{31}, τ_{12}, τ_{22}	*Intercepts* associés aux variables manifestes x reflétant les variables latentes endogènes
τ_1 à τ_5	*Intercepts* associés aux variables manifestes y reflétant les variables latentes exogènes
τ_{ξ_1} à τ_{ξ_2}	*Intercepts* associés aux variables latentes endogènes
λ_{11}, λ_{21}, λ_{31}, λ_{12}, λ_{22}	Coefficients (*factor loadings*) reliant les variables manifestes x aux variables latentes endogènes
λ_1 à λ_5	Coefficients (*factor loadings*) reliant les variables manifestes y à la variable latente exogène
β_{12}, γ_1 et γ_2	Coefficients structurels
ϵ_{11}, ϵ_{21}, ϵ_{31}, ϵ_{12}, ϵ_{22}	Termes d'erreurs (résidus) associées aux variables manifestes x
δ_1 à δ_5	Termes d'erreurs (résidus) associées aux variables manifestes y
ζ_1 et ζ_2	Résidus latents associés aux variables latentes endogènes (termes de pertubation latente)

TABLEAU 2.1 – Notations du SEM considéré

Les trois premiers systèmes d'équations (2.1) à (2.3) décrivent les liens entre les variables manifestes et les variables latentes, tandis que le dernier système d'équations décrit les liens entre les variables latentes endogènes (ξ_1 et ξ_2) et la variable latente exogène (η). On peut remarquer que les SEM supposent des liens linéaires entre les variables manifestes et variables latentes. Il en va de même pour les liens entre les variables latentes. Les hypothèses associées au SEM considéré sont :

- Les résidus des modèles de mesure (ε et δ), les résidus latents du modèle structurel (ζ), et les variables latentes (ξ_1, ξ_2 et η) sont centrés et normalement distribués.
- les résidus ε des modèles de mesure pour les variables latentes endogènes sont non corrélés avec les résidus δ du modèle de mesure pour la variable exogène.
- Les résidus des modèles de mesure (ε et δ) sont non corrélés avec les variables latentes (ξ_1, ξ_2 et η) et les résidus latents ζ.
- Les résidus latents du modèle structurel (ζ) sont non corrélés avec la variable latente exogène (η).

Dans le modèle de la figure 2.1, on a également supposé que les termes d'erreurs d'un même modèle de mesure n'étaient pas corrélés entre eux. Cette hypothèse est néanmoins parfois relâchée. Pour ce faire, on ajoute une flèche bidirectionnelle entre les résidus que l'on souhaite considérer corrélés : cette flèche indique qu'un nouveau coefficient est à estimer (c.-à-d., la corrélation entre les deux résidus). Idem, on a supposé que les résidus latents n'étaient pas corrélés.

2.2.4 Estimation des paramètres

L'élément essentiel des modèles à équations structurelles est la covariance entre les variables manifestes. Ces covariances sont souvent regroupées dans une matrice de dimension $p \times p$ (où p désigne le nombre de variables manifestes) appelée matrice de covariance :

$$\begin{pmatrix} \mathrm{Var}\,(X_1) & \mathrm{Cov}\,(X_1, X_2) & \cdots & \mathrm{Cov}\,(X_1, X_p) \\ \mathrm{Cov}\,(X_2, X_1) & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}\,(X_p, X_1) & \cdots & \cdots & \mathrm{Var}\,(X_p) \end{pmatrix}$$

Si la "vraie" matrice de covariance en population Σ n'est pas accessible, il est néanmoins possible de l'estimer à l'aide des données recueillies. On parle alors de matrice de covariance observée sur l'échantillon (matrice S). Par ailleurs, à partir des paramètres d'un SEM, il est possible d'exprimer la matrice de covariance "attendue" entre les variables manifestes. Cette matrice s'écrit donc en fonction des paramètres du modèle et se note $\Sigma(\theta)$.

Le processus d'estimation d'un SEM vise à trouver les valeurs des paramètres du modèle (notées $\hat{\theta}$) qui permettent d'obtenir la matrice $\Sigma(\hat{\theta})$ la plus proche possible de la matrice S. Autrement dit, on cherche à trouver les estimations $\hat{\theta}$ pour lesquelles l'écart entre S et $\Sigma(\theta)$ est minimal. Une fonction possible pour caractériser l'écart entre ces deux matrices est la fonction du maximum de vraisemblance [27] :

$$F_{MV} = \ln|\Sigma(\theta)| - \ln|S| + \mathrm{tr}(S \times (\Sigma(\theta))^{-1}) - p \tag{2.5}$$

Où :

– ln désigne la fonction logarithme népérien

– $|\Sigma(\theta)|$ et $|S|$ désignent respectivement les déterminants des matrices $\Sigma(\theta)$ et S. Le déterminant est un nombre. Il s'agit d'une mesure généralisée de la variance de l'ensemble des variables contenues dans la matrice. Si les deux matrices sont identiques, alors leurs déterminants le seront aussi.

– Le terme $\text{tr}(S \times (\Sigma(\theta))^{-1})$ désigne la trace du produit entre la matrice S et la matrice inverse de $\Sigma(\theta)$ (notée $(\Sigma(\theta))^{-1}$). La trace d'une matrice correspond à la somme de ses coefficients diagonaux. Ainsi, si les matrices S et $\Sigma(\theta)$ sont identiques, le produit $S \times (\Sigma(\theta))^{-1}$ correspondra à la matrice identité I_p et la trace de ce produit vaudra p (p étant le nombre de variables manifestes).

Lorsqu'un modèle s'ajuste parfaitement aux données, la valeur de la fonction d'ajustement est égale à zéro. Par ailleurs, sous l'hypothèse nulle $H_0 : \Sigma(\theta) = \Sigma$, la statistique de test $\chi^2 = (N-1) \times F_{MV_{\min}}$ suit une loi du chi-deux à $p(p+1)/2 - q$ degrés de liberté (où N désigne l'effectif de l'échantillon, $F_{MV_{min}}$ est la variable aléatoire "minimum de la fonction F_{MV}" dont la valeur peut varier en fonction de l'échantillon considéré, p est le nombre de variables manifestes et q est le nombre de paramètres à estimer [2]).

Cette méthode d'estimation visant à minimiser l'écart entre S et $\Sigma(\theta)$ en se basant sur la fonction F_{MV} est appelée estimation par maximum de vraisemblance. Elle suppose la multinormalité des variables manifestes. Aussi, les variables manifestes doivent donc être continues. En pratique, l'utilisation de l'estimation par maximum de vraisemblance est néanmoins fréquente pour des variables manifestes catégorielles ordinales. En cas de nombre de modalités inférieur à sept, une correction de Sattora-Bentler est souvent appliquée [28]. Cette méthode permet de corriger la statistique de test χ^2 et les erreurs standards du modèle afin de prendre en compte le niveau de non-normalité des variables [29].

2. Pour que le modèle soit identifiable, il faut que le nombre de paramètres à estimer q soit inférieur ou égal à $p(p+1)/2$ où p est le nombre de variables manifestes. Le nombre $p(p+1)/2$ correspond aux valeurs uniques de la matrice de covariance des variables manifestes.

Lorsque les variables manifestes sont ordinales et qu'elles présentent un nombre de modalités faible (par exemple inférieur ou égal à 4), une autre méthode d'estimation plébiscitée est la méthode *Diagonally Weighted Least Squares* [30, 31]. Avec cette méthode, les variables manifestes ordinales x sont supposées être une simplification d'une variable continue sous-jacente distribuée normalement x^* (parfois appelée "réponse latente"). Les variables observés x seraient alors le résultat de la troncature de la réponse latente x^* (par l'intermédiaire de seuils s, voir figure 2.2). C'est à partir des corrélations entre ces nouvelles variables (corrélations polychoriques) que l'estimation des paramètres du modèle a ensuite lieu.

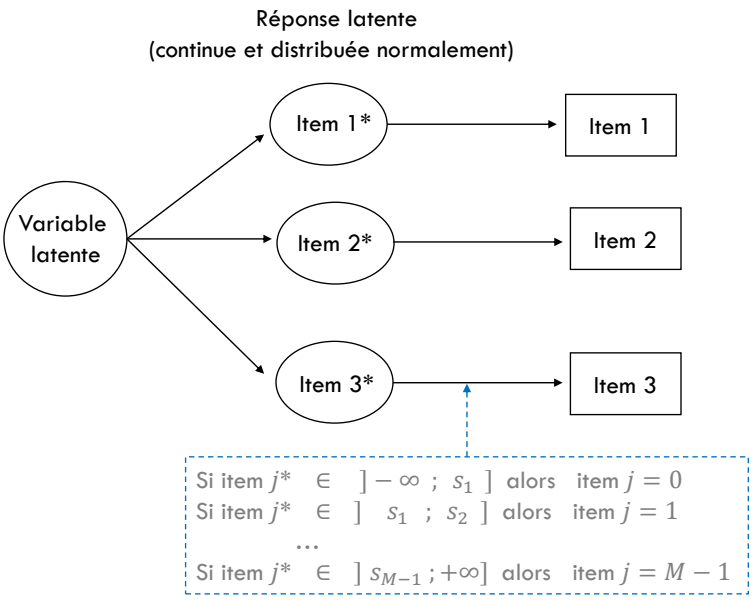


FIGURE 2.2 – Modèle à équations structurelles où trois items ordinaux (ayant M modalités de réponse) sont supposés être une simplification d'une réponse latente

2.2.5 Évaluation de l'ajustement du modèle

Pour pouvoir évaluer l'adéquation (ou l'ajustement) du modèle testé, de nombreux indices ont été proposés (indices de *fit*). Comme chacun de ces indices reflète un aspect particulier de l'ajustement, aucun indice ne devrait être utilisé seul. Kline [32] a indiqué en 2016 qu'il fallait au minimum rapporter :

– **La réalisation de la statistique de test du chi-deux et son degré de liberté** :

Réalisation sur l'échantillon : $\chi^2_{exp} = (N-1) \times f_{MV_{\min}}$

Degré de liberté : $ddl = p(p+1)/2 - q$

Dans la littérature, on retrouve parfois chi-deux normé qui correspond au rapport entre ses deux quantités. Le chi-deux normé est attendu aussi petit que possible.

– ***Root-Mean-Square Error of approximation*** **(RMSEA)** [33–35] :

Le RMSEA est un indice d'ajustement absolu évaluant directement la correspondance entre le modèle spécifié et les données observées. Il est défini comme suit :

$$RMSEA = \sqrt{\frac{\max\left(\chi^2_{exp} - ddl, 0\right)}{(N-1) \times ddl}}$$

Le terme $\max\left(\chi^2_{exp} - ddl, 0\right)$ est appelé la mesure de non-centralité. On souhaite cette mesure aussi petite que possible.

– ***Standardized Root-Mean-square Residual*** **(SRMR)** [36] :

Comme le RMSEA, le SRMR est un indice d'ajustement absolu. Il s'intéresse à la différence entre les corrélations observées entre les variables manifestes et les corrélations prédites par le modèle.

$$SRMR = \sqrt{\frac{1}{\frac{p(p+1)}{2}} \sum_{j=1}^{p} \sum_{j'=1}^{j} \left(\frac{s_{jj'} - \widehat{\sigma}_{jj'}}{\sqrt{s_{jj} s_{j'j'}}} \right)^2}$$

Avec :

- $\frac{p(p+1)}{2}$: Le nombre de paramètres uniques dans la matrice de covariance des variables manifestes (les valeurs sur la diagonale et en dessous)
- $s_{jj'}$: la covariance observée entre les variables manifestes indicées par j et j'
- $\hat{\sigma}_{jj'}$: la covariance prédite (ou induite) par le modèle entre les variables manifestes indicées par j et j'
- s_{jj} et $s_{j'j'}$: les variances observées des variables manifestes indicées par j et j', respectivement

– ***Comparative Fit Index* (CFI)** [37] :

Le CFI est un indice d'ajustement incrémentiel. Il compare l'ajustement du modèle étudié à celui d'un modèle "nul" qui ne supposerait aucune relation entre les variables manifestes. La valeur du chi-deux et le degré de liberté de ce modèle "nul" sont notés respectivement χ^2_{nul} et ddl_{nul}. Le CFI est alors défini par :

$$CFI = \frac{\max\left(\chi^2_{nul} - ddl_{nul}, 0\right) - \max\left(\chi^2_{exp} - ddl, 0\right)}{\max\left(\chi^2_{nul} - ddl_{nul}, 0\right)}$$

Cet indice traduit l'amélioration de la mesure de non-centralité entre le modèle nul et le modèle étudié. On souhaite que cet indice soit le plus grand possible.

En pratique, les valeurs de ces indices sont comparées à des valeurs seuils afin de définir si l'adéquation du modèle est bonne, acceptable, médiocre ou mauvaise [38]. Pour chacun de ces indices, tous les auteurs ne s'accordent pas sur les valeurs de seuils définissant un ajustement bon ou acceptable. On retrouve donc dans la littérature une certaine variabilité des seuils utilisés. Ces comparaisons à des valeurs seuil restent par ailleurs des règles arbitraires.

2.2.6 Re-spécification du modèle

En cas de mauvais ajustement du modèle, il est possible de le re-spécifier. La re-spécification du modèle a pour but d'améliorer l'ajustement du modèle en supprimant (ou en ajoutant) des contraintes sur certains paramètres. Par exemple, les corrélations entre les résidus d'un modèle

de mesure sont souvent contraintes à être nulles. Néanmoins, il est possible que cette contrainte soit trop stricte : dans ce cas, on va chercher à identifier les corrélations résiduelles qu'il pourrait être intéressant de libérer et on estimera ensuite un nouveau modèle où elles seront estimées. L'ajustement du modèle ainsi re-spécifié sera comparé à celui du modèle initial pour déterminer si cette re-spécification a été bénéfique ou non.

Lorsque le modèle ne s'ajuste pas correctement aux données, il est fréquent de se tourner vers les indices de modifications. Ces indices sont des outils utilisés pour identifier les modifications qui pourraient conduire à un meilleur ajustement du modèle. Ils sont associés aux paramètres fixes (ou contraints) du modèle et donnent une approximation de la diminution attendue du χ^2 du modèle si le paramètre était libéré. Une valeur élevée pour un indice de modification indique que la libération du paramètre correspondant pourrait substantiellement améliorer l'ajustement du modèle. Une stratégie possible consiste alors à libérer un à un les paramètres fixes/contraints associés à des indices de modification élevés (en commençant par le paramètre associé à l'indice de modification le plus élevé).

L'apport des modifications faites au modèle peut être testé grâce à un test du rapport de vraisemblance. Comme son nom l'indique, ce test permet de comparer deux modèles A et B en calculant le rapport de leur vraisemblance. Sous l'hypothèse nulle de vraisemblances identiques, la statistique de test suit une loi du chi-deux dont le degré de liberté correspond à la différence de degrés de libertés entre les deux modèles. Un test significatif s'interprète comme une amélioration de l'ajustement du modèle. Pour que ce test soit réalisable, les deux modèles doivent être emboîtés et estimés à partir du même ensemble de données.

Les modifications apportées au modèle ne doivent pas être réalisées dans le seul objectif d'améliorer l'ajustement du modèle. Elles doivent avoir un fondement théorique ou empirique et doivent pouvoir être interprétées. La re-spécification du modèle doit être envisagée comme une étape permettant de corriger le modèle en réalisant des modifications qui restent cohérentes avec le modèle initialement spécifié.

2.2.7 SEM et validité de structure d'un questionnaire

Lors du développement d'un questionnaire, les liens unissant les variables latentes (représentant les dimensions du questionnaire) et les variables manifestes (les items) doivent être définis en amont, sur des fondements théoriques. On parle de "structure du questionnaire", définie *a priori*. La structure ainsi définie doit ensuite être éprouvée pour déterminer si elle est valide ou non.

Dans le cadre de la théorie classique des tests, lorsque l'on cherche à valider la structure d'un questionnaire, on a souvent recours aux modèles à équations structurelles. Concrètement, un SEM est estimé à partir de données empiriques (passation du questionnaire par des individus). Les liens de ce SEM doivent correspondre à ceux décrits dans la structure *a priori* du questionnaire : on parle alors d'analyse factorielle confirmatoire (*Confirmatory Factor Analysis*, CFA). Lors d'une CFA, on ne suppose généralement pas de relations causales entre les variables latentes. Au lieu de cela, on modélise simplement les liens qui les unissent en estimant leur covariance. Toutefois, on peut également considérer que les corrélations entre les variables latentes sont expliquées par une variable latente de second ordre (*higher order factor*). On spécifie alors un modèle structurel où une variable latente "globale" est reflétée par les variables latentes de premier ordre (représentant les dimensions). On parle dans ce cas de CFA hiérarchique (voir figure 2.3).

Pour qu'une structure soit valide, le modèle doit présenter un bon ajustement (ou tout du moins un ajustement acceptable), évalué grâce aux indices d'ajustement présentés dans la section précédente. De plus, les estimations des *factor loadings* reliant les items à leur dimension hypothétique doivent être élevées (supérieurs à 0,4 ou 0,5).

La même stratégie est utilisée lorsque l'on cherche à valider la structure d'un questionnaire nouvellement traduit. La structure du SEM utilisé pour réaliser l'analyse factorielle confirmatoire correspond alors à celle proposée par les auteurs du questionnaire princeps.

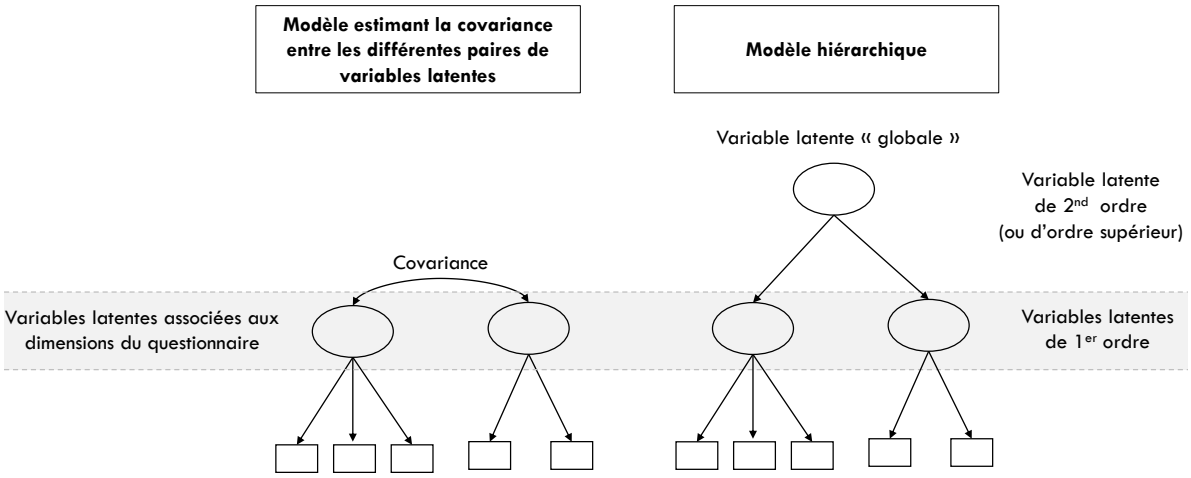


FIGURE 2.3 – Analyse factorielle confirmatoire et analyse factorielle confirmatoire hiérarchique

2.3 Théorie de la réponse aux items

Notations : Pour toute cette section, on supposera disposer d'un ensemble de J items (indicés par $j = 1, ..., J$) et d'un échantillon composé de N individus (indicés par $i = 1, ..., N$).

La théorie de la réponse aux items est une famille de modèles qui se concentrent sur l'information au niveau des items. Ces modèles cherchent à caractériser le lien entre la variable latente que l'on cherche à mesurer et les réponses observées aux items visant à la mesurer.

Dans le cadre de la théorie de la réponse aux items, on ne modélise pas directement la réponse à l'item. À la place, on modélise, pour chaque item, la probabilité de choisir une certaine modalité de réponse. Ces probabilités sont modélisées en fonction des caractéristiques des individus (c'est-à-dire leur niveau individuel de variable latente θ) et des items. Mathématiquement, la probabilité que l'individu i choisisse la modalité de réponse x à l'item j, s'écrit sous la forme d'une fonction à plusieurs paramètres, appelée fonction de réponse à l'item (*Item Reponse Function*, IRF) :

$$P(X_{ij} = x \mid \theta_i, \boldsymbol{\psi_j}) = f(\theta_i, \boldsymbol{\psi_j}) \tag{2.6}$$

La fonction de réponse à l'item est une probabilité conditionnelle : elle dépend des caractéristiques de l'item j (regroupées dans le vecteur $\boldsymbol{\psi_j}$) et du niveau de variable latente de l'individu i considéré (θ_i). Pour un item dichotomique, on ne présente généralement que la probabilité de répondre 1 à l'item (la probabilité de l'événement complémentaire "répondre 0" en découlant directement).

Les premiers modèles de la théorie de la réponse aux items sont apparus dans la seconde partie du 20^{e} siècle et étaient initialement conçus pour l'analyse d'items dichotomiques (modalités 0 = "non" et 1 = "oui"). Le premier modèle que l'on peut relier à la théorie de la réponse aux items est le modèle déterministe de Guttman [39], bien qu'il lui soit en fait légèrement antérieur. Avec son modèle, Guttman suppose qu'il existe un lien déterministe entre le niveau de la variable latente et la réponse à un item donné. Concrètement, il suppose l'existence d'un niveau "seuil" de variable latente au-delà duquel les individus répondent forcément 1 (probabilité de répondre 1 égale à 100%) et en deçà duquel les individus répondent toujours 0 (probabilité de répondre

1 égale à 0%). Quelques années plus tard, deux modèles probabilistes pour items dichotomiques sont apparus :

- Le modèle de Rasch, développé par le mathématicien Georg Rasch afin d'obtenir une mesure du construit étudié indépendante de l'instrument de mesure [40].
- Le modèle logistique à deux paramètres (2-PLM), développé par Birnbaum [41] et correspondant à la vision de la psychométrie de Lord (c.-à-d., qui vise à obtenir un modèle qui s'ajuste bien aux données) [42]

Nous verrons dans les sections suivantes que ces modèles sont mathématiquement proches. Néanmoins, ils sont basés sur deux visions de la psychométrie différentes. La vision de Rasch est centrée sur le modèle. Ainsi, pour obtenir une mesure objective du construit, Rasch cherche à sélectionner des items qui satisfont son modèle. La vision de Lord est, au contraire, centrée sur les données. L'objectif est alors de proposer des modèles qui s'adaptent au mieux aux données recueillies, grâce à une caractérisation fine des items (permise par l'introduction de différents paramètres). Par la suite, d'autres modèles ont fait leur apparition pour caractériser toujours plus finement les items ou pour pouvoir considérer des items polytomiques.

Le modèle de Rasch et ses extensions pour items polytomiques ont longtemps été regroupés avec les modèles issus de la famille de Lord sous le nom de théorie de la réponse aux item (*Item Response Theory*, IRT). Ces deux familles de modèles ont cependant récemment formalisé leur séparation [43, 44]. Aujourd'hui, les modèles de la famille de Rasch forment donc leur propre famille : la théorie de la mesure de Rasch (*Rasch Measurement Theory*, RMT). Les modèles visant à caractériser finement les items sont, quant à eux, toujours désignés sous le terme de théorie de la réponse aux items (on parle aussi de modèles de la famille de Lord).

Les sections suivantes présentent en détail les trois modèles mentionnés ci-dessus, ainsi que certains modèles qui en ont découlé et auxquels ce manuscrit se réfère. Enfin, les hypothèses fondamentales de ces familles de modèles seront rappelées.

2.3.1 Prémices de l'IRT : le modèle déterministe de Guttman

Modèle de Guttman pour items dichotomiques

Le modèle de Guttman pour items dichotomiques (modalités de réponse : 0 = "non" / 1 = "oui") modélise de façon déterministe la probabilité de répondre favorablement à un item en fonction du niveau de la variable latente de l'individu θ_i. Ce modèle suppose qu'il existe, pour chaque item dichotomique, un niveau de variable latente "seuil" δ_j tel que :

- Les individus ayant un niveau de variable latente inférieur à ce seuil répondent toujours 0 (ils "ne réussissent pas" l'item, ou répondent "défavorablement" à l'item)
- Les individus ayant un niveau de variable latente supérieur à ce seuil répondent toujours 1 (ils "réussissent" l'item, ou répondent "favorablement" à l'item).

La fonction réponse à l'item IRF associée à ce modèle s'écrit alors pour tout item j :

$$P(X_{ij} = 1 \mid \theta_i, \delta_j) = \begin{cases} 0 \text{ si } \theta_i < \delta_j \\ 1 \text{ si } \theta_i > \delta_j \end{cases} \tag{2.7}$$

Avec X_{ij} la réponse de l'individu i à l'item j, θ_i le niveau de variable latente de l'individu i et δ_j le paramètre de seuil caractérisant l'item. La représentation graphique de cette fonction, appelée courbe caractéristique de l'item (*Item Caracteristic Curve*, ICC), est donnée en figure 2.4.

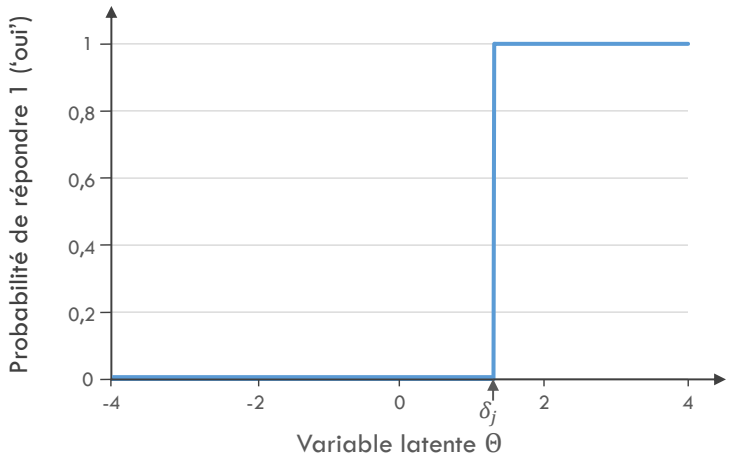


FIGURE 2.4 – Courbe caractéristique d'un item j dichotomique vérifiant le modèle de Guttman

Dans le cadre du modèle de Guttman, on cherche à ordonner les items du plus "facile" (item le plus populaire) au plus "difficile" (item le moins populaire) en calculant leur difficulté d_j. Dans ce manuscrit, la difficulté d'un item est définie (dans le cadre du modèle de Guttman) comme la proportion d'individus ayant répondu négativement à l'item. À partir des données d'un échantillon de taille N, la difficulté d_j de l'item j est définie par :

$$d_j = \frac{\#(X_{ij} = 0)}{N} \tag{2.8}$$

Si cette proportion est faible, alors l'item est réussi par beaucoup d'individus : il est considéré comme facile (ou populaire). Au contraire, si la proportion d'individus ayant répondu négativement à l'item est élevée, alors l'item a été peu réussi : on dit qu'il est difficile (ou non populaire). La notion de difficulté est directement liée au paramètre de seuil : un item sera d'autant plus difficile que son paramètre de seuil est élevé et inversement (voir figure 2.5, pour une représentation simultanée des courbes caractéristiques de trois items ayant des paramètres de seuils différents). Par exemple, si l'on considère les trois items dichotomiques suivants (supposés mesurer la dépression) : item 1 = "Je me sens triste", item 2 = "Je me sens déprimé(e)" et item 3 = "J'ai l'impression de n'avoir aucune raison de vivre". On s'attend à ce que la difficulté de l'item 3 soit plus élevée que celle de l'item 2, elle-même plus élevée que celle de l'item 1.

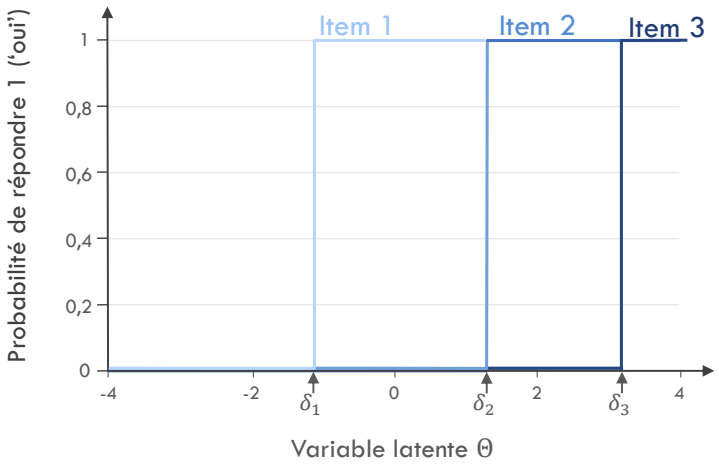


FIGURE 2.5 – Courbes caractéristiques de trois items dichotomiques vérifiant le modèle de Guttman

Avec le modèle de Guttman, le nombre de profils de réponse qu'il est possible d'observer est très restreint (puisque si un item est réussi, alors tous les items plus faciles doivent aussi l'être). Ainsi, si l'on considère les trois items dichotomiques de la figure 2.5, on ne devrait pouvoir observer que les quatre profils de réponse suivants :

	Item 1 : $\boldsymbol{X_1}$	**Item 2 :** $\boldsymbol{X_2}$	**Item 3 :** $\boldsymbol{X_3}$
Profil 1	0	0	0
Profil 2	1	0	0
Profil 3	1	1	0
Profil 4	1	1	1

La terminologie utilisée dans les paragraphes précédents ("réussir l'item", "item facile", "item difficile", "difficulté") est empruntée aux sciences de l'éducation. En effet, c'est de ce domaine que sont issus les modèles de réponse à l'item. Dans le champ de la santé, cette terminologie peut parfois paraître peu adéquate, mais elle reste employée de façon courante.

Modèle de Guttman pour items polytomiques

Le modèle de Guttman peut être étendu aux items polytomiques. Le principe de la modélisation pour les items polytomiques est décrit ci-dessous, en considérant un item j ayant M_j modalités de réponse : 0, 1, 2, ..., $M_j - 1$.

Pour chaque modalité de réponse p strictement supérieure à 0 de l'item j, on peut définir une variable binaire I_{jp} de sorte que :

$$I_{jp} = \mathbb{1}_{\{\text{item } j \geq p\}} = \begin{cases} 0 \text{ si item } j < p \\ 1 \text{ si item } j \geq p \end{cases} \tag{2.9}$$

La nouvelle variable I_{jp} indique, pour chaque individu, si la modalité de réponse p de l'item j est atteinte ou non. On définit au total $M_j - 1$ variables (autant que de modalités de réponse strictement supérieures à 0). On ne définit pas de variable I_{jp} pour la modalité 0, car cette dernière est toujours atteinte (un individu répondant toujours 0 ou une modalité supérieure si, bien sûr, il choisit de répondre à l'item). Ces variables sont parfois appelées *item-step*.

Par exemple, si l'on considère l'item j = "Vous sentez-vous triste ?" (modalités : 0 = "pas du tout", 1 = "un peu", 2 = "beaucoup" et 3 = "totalement"). La variable I_{j1} indique, pour chaque individu, si la modalité "un peu" est atteinte. Ce sera le cas si et seulement si l'individu a répondu "un peu", "beaucoup" ou "totalement". La variable I_{j2} indique, quant à elle, si la modalité "beaucoup" est atteinte, ce sera le cas si et seulement si l'individu a répondu "beaucoup" ou "totalement". On définit de façon similaire la variable I_{j3}.

Comme précédemment, le modèle de Guttman suppose qu'il existe, pour chaque variable I_{jp} (chaque *item-step*), un niveau de variable latente "seuil" δ_{jp} tel que :

- Les individus ayant un niveau de variable latente inférieur à ce seuil vérifient $I_{jp} = 0$. Autrement dit, ces individus n'atteignent pas la modalité de réponse p, leur réponse à l'item j est donc inférieure à p : $X_j < p$.
- Les individus ayant un niveau de variable latente supérieur à ce seuil vérifient $I_{jp} = 1$. En d'autres termes, ils atteignent la modalité de réponse p. Leur réponse à l'item j est donc supérieure ou égale à p : $X_j \geq p$

Les fonctions réponse des *item-steps* I_{jp} de l'item j s'écrivent alors :

$$P(I_{jp_i} = 1 \mid \theta_i, \delta_{jp}) = \begin{cases} 0 \text{ si } \theta_i < \delta_{jp} \\ 1 \text{ si } \theta_i > \delta_{jp} \end{cases} \tag{2.10}$$

Autrement dit, pour la réponse à l'item j on obtient :

$$P(X_{ij} \geq p \mid \theta_i, \delta_{jp}) = \begin{cases} 0 \text{ si } \theta_i < \delta_{jp} \\ 1 \text{ si } \theta_i > \delta_{jp} \end{cases} \tag{2.11}$$

Les représentations graphiques de ces fonctions sont données en figure 2.6 avec un item à quatre modalités de réponse (0, 1, 2, 3) vérifiant le modèle de Guttman.

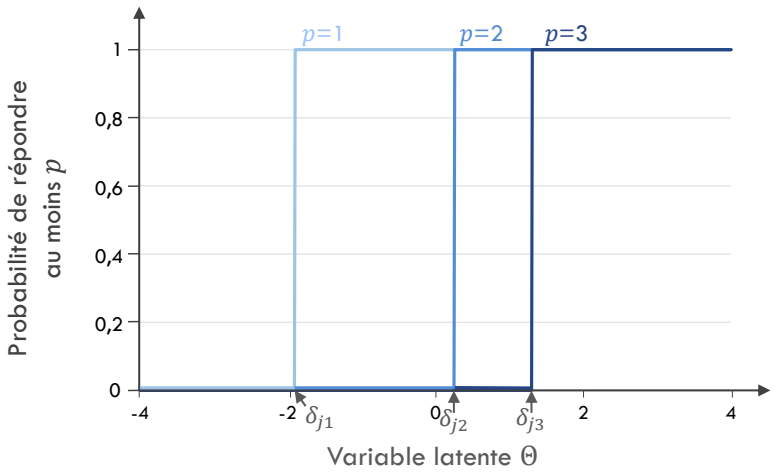


FIGURE 2.6 – Représentation graphique des fonctions de réponse des *item-steps* I_{jp} associées à un item j ayant quatre modalités de réponse ($p = 0, 1, 2, 3$) vérifiant le modèle de Guttman

À partir de l'équation (2.11), on peut associer à chaque niveau de variable latente le score à l'item j attendu d'après le modèle de Guttman (voir figure 2.7 pour une représentation graphique) :

$$X_{ij} = \begin{cases} 0 \text{ si } \theta_i < \delta_{j1} \\ 1 \text{ si } \delta_{j1} < \theta_i < \delta_{j2} \\ \dots \\ M_j - 1 \text{ si } \theta_i > \delta_{jM_j-1} \end{cases} \tag{2.12}$$

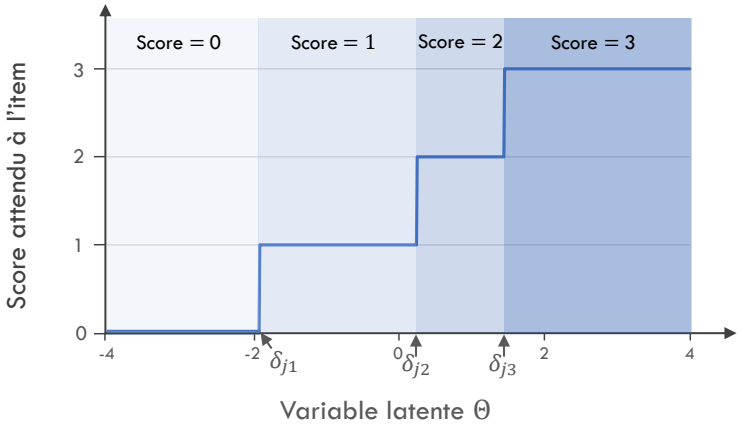


FIGURE 2.7 – Score attendu d'après le modèle de Guttman pour un item j à quatre modalités de réponse (0, 1, 2, 3)

En présence de plusieurs items polytomiques, on cherche à ordonner les modalités de réponse de la plus "facile à atteindre" à la plus "difficile". Pour ce faire, on calcule leur difficulté. Dans ce manuscrit, la difficulté de la modalité p de l'item j correspond à la proportion d'individus n'ayant pas atteint cette modalité. À partir des données d'un échantillon de taille N, la difficulté d_{jp} de la modalité p l'item j est définie par :

$$d_{jp} = \frac{\#(X_{ij} < p)}{N} \tag{2.13}$$

Si la difficulté est faible, cela signifie que la modalité a été atteinte par beaucoup d'individus : on la considèrera comme "facile à atteindre". Au contraire, si la difficulté est élevée, cela signifie que peu d'individu ont réussi à l'atteindre. Dans ce cas, on la considèrera "difficile". Les difficultés ne sont définies que pour les modalités p strictement positives. La difficulté d_{jp} de la modalité de réponse p est directement liée au paramètre de seuil δ_{jp} associé à l'*item-step* I_{jp}. Une modalité sera d'autant plus difficile que son paramètre de seuil est élevé (et inversement).

Comme avec les items dichotomiques, le nombre de profils de réponse qu'il est possible d'observer est restreint avec le modèle de Guttman pour les items polytomiques. En effet, si une modalité de réponse est atteinte, alors toutes les modalités de réponse plus faciles doivent également l'être.

Les erreurs de Guttman

Le modèle de Guttman, de par sa nature déterministe, n'est pas réaliste. Il est néanmoins parfois utilisé lors de l'analyse de données rapportées les patients. On l'utilise alors comme modèle de référence, et l'on cherche à déterminer à quel point le modèle sous-jacent aux données recueillies s'en écarte. Afin de quantifier cet écart, on utilise les erreurs de Guttman :

- **Pour des items dichotomiques :** une erreur de Guttman survient dans les réponses d'un individu dès que ce dernier réussit un item, alors qu'il échoue à un item plus facile.
- **Pour des items polytomiques :** une erreur de Guttman survient dans les réponses d'un individu dès qu'il atteint une modalité de réponse, alors qu'il n'a pas atteint une autre modalité plus facile.

On parle d'erreurs de Guttman au sens de "déviations par rapport au modèle de Guttman" (on parle aussi parfois d'incohérences au sens de Guttman). Cette terminologie est à prendre avec précaution : il n'y a, bien sûr, pas d'"erreurs" dans les réponses des individus.

Le nombre d'erreurs de Guttman au sein des réponses d'un individu permet de quantifier le degré de cohérence de ses réponses (dans la littérature, on parle d'indice de *person fit* non paramétrique [45]). Un faible nombre d'erreurs de Guttman indiquera que les réponses de l'individu sont assez cohérentes, par exemple (avec des items dichotomiques) :

- L'individu a réussi les items faciles, mais n'a pas réussi les items difficiles ;
- L'individu a à la fois réussi les items faciles et les items difficiles ;
- L'individu n'a réussi ni les items faciles ni les items difficiles.

En revanche, un nombre important d'erreurs de Guttman indiquera des réponses peu cohérentes (c.-à-d. un individu qui réussit les items difficiles, mais qui ne réussit pas les items faciles).

La méthode de calcul du nombre d'erreurs de Guttman pour un individu donné est réexplicitée dans le chapitre 4 dans le cadre d'items polytomiques.

2.3.2 Les modèles de la famille de Rasch

Au cours des années 1950, le mathématicien Georg Rasch a cherché un modèle qui permette d'obtenir une mesure du construit étudié indépendante de l'instrument de mesure. Il a proposé, en 1960, un modèle pour items dichotomiques : le modèle de Rasch [40]. Suite à ces travaux, deux extensions pour items polytomiques ont été proposées : le modèle de crédit partiel [46] et le *rating scale model* [47–49]. Les paragraphes suivants présentent en détail le modèle de Rasch pour item dichotomiques et le modèle de crédit partiel.

Le modèle de Rasch pour items dichotomiques

Ce modèle a été développé par Rasch en 1960 [46] pour obtenir une mesure objective du construit étudié (c'est-à-dire une mesure qui ne dépend pas de l'instrument de mesure). Le modèle de Rasch est stochastique (par opposition au modèle de Guttman) et est conçu pour des items dichotomiques (modalités : 0 = "non" et 1 = "oui"). Avec ce modèle, la fonction de réponse d'un item dichotomique j a la forme d'un modèle logistique. Elle est donnée par :

$$P(X_{ij} = 1 \mid \theta_i, \delta_j) = \frac{\exp(\theta_i - \delta_j)}{1 + \exp(\theta_i - \delta_j)} \tag{2.14}$$

Avec X_{ij} la réponse de l'individu i à l'item j, θ_i le niveau de variable latente de l'individu i et δ_j l'unique paramètre caractérisant l'item. Le paramètre δ_j est communément appelé "paramètre de seuil" de l'item j (*item threshold*). Dans la littérature, on retrouve également les termes *item difficulty* ou *item location*. Ce paramètre représente le niveau de variable latente pour lequel la probabilité de répondre 1 à l'item vaut 50% (c.-à-d. le niveau de variable latente pour lequel il y a autant de chance de réussir l'item que de ne pas le réussir). Plus le paramètre de seuil d'un item est élevé, plus l'item associé est considéré difficile. La courbe caractéristique d'un item dichotomique vérifiant le modèle de Rasch est donnée en figure 2.8.

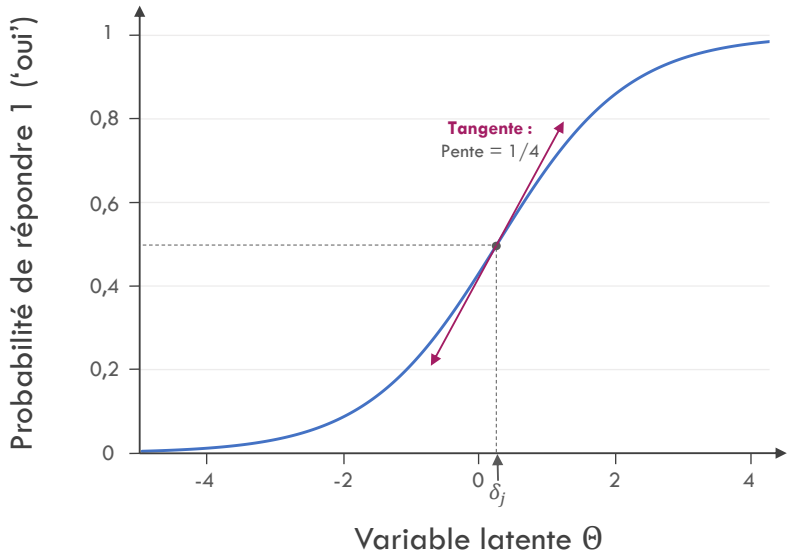


FIGURE 2.8 – Courbe caractéristique d'un item j dichotomique vérifiant le modèle de Rasch

On peut voir que :

- Plus le niveau de variable latente de l'individu i est élevé, plus la probabilité que l'individu réponde 1 à l'item est grande.
- Pour des niveaux de variable latente très faibles (vers $-\infty$), la probabilité de répondre 1 tend vers 0 (asymptote horizontale à gauche).
- Pour des niveaux de variable latente très grands (vers $+\infty$), la probabilité de répondre 1 tend vers 1 (asymptote horizontale à droite).
- La pente de la tangente à la courbe au point d'inflexion (au niveau du paramètre de seuil de l'item) vaut 1/4. Ce sera toujours le cas avec le modèle de Rasch. En conséquence, les ICC des items vérifiant le modèle de Rasch ne se coupent jamais : on dit qu'elles sont "doublement monotone". Dans le cas précis du modèle de Rasch, on dit même que ces courbes sont parallèles (voir figure 2.9 pour une illustration avec trois items).

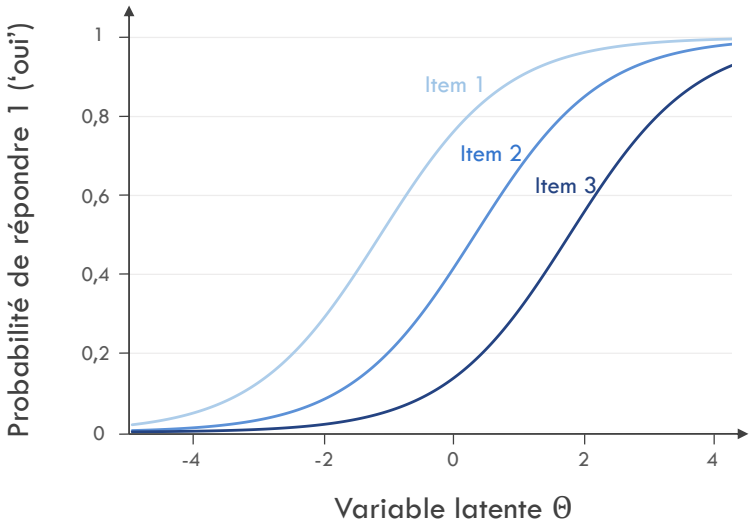


FIGURE 2.9 – Courbes caractéristiques de trois items vérifiant le modèle de Rasch

Ce modèle présente des propriétés psychométriques spécifiques d'intérêt pour l'analyse des données autorapportées :

- **L'exhaustivité du score sur la variable latente :** Le score total de l'individu (calculé comme la somme de ses réponses sur l'ensemble des items : $S_i = \sum_{j=1}^{J} X_{ij}$) contient toute l'information nécessaire pour déterminer son niveau individuel de variable latente. À chaque valeur de score observée correspond une unique valeur de variable latente. Par conséquent, tous les individus ayant le même score auront la même estimation pour le niveau de la variable latente (indépendamment du profil de réponse permettant d'obtenir ce score). En l'absence de données manquantes, le score observé peut être utilisé pour ordonner les individus par niveau de variable latente.

- **L'objectivité spécifique :** La mesure obtenue grâce au modèle de Rasch est indépendante de l'instrument de mesure [50]. Ainsi, les estimations du niveau de variable latente des individus (θ_i) ne dépendent pas des paramètres de seuil des items utilisés (δ_j). Réciproquement, les estimations des paramètres de seuil des items ne dépendent pas de l'échantillon. Cette propriété est intéressante en cas de données manquantes. En effet, on pourrait estimer sans biais le niveau individuel de variable latente des individus ayant des questionnaires incomplets (individus n'ayant pas répondu à tous les items), et ce même en cas de données manquantes de type MNAR (*Missing Not At Random* [51]).

Le modèle de crédit partiel pour items polytomiques

Deux extensions du modèle de Rasch (bénéficiant des mêmes propriétés) ont été proposées pour pouvoir considérer des items à plus de deux modalités de réponse (items polytomiques) : le modèle de crédit partiel (*Partial Credit Model*, PCM) [46] et le *Rating Scale Model* (RSM) [47–49]. Le RSM pouvant être vu comme un cas particulier du PCM, cette section est limitée à ce dernier modèle.

Le fait de pouvoir considérer des items polytomiques est particulièrement intéressant dans le cadre des données rapportées par les patients. En effet, les échelles de réponse des questionnaires utilisés en santé sont souvent ordinales (par exemple : "pas du tout", "un peu", "beaucoup", "totalement" ou "jamais","rarement", "quelques fois", "souvent", "très souvent", "en permanence").

Considérons un ensemble de J items polytomiques ayant chacun M_j modalités de réponse. Le PCM donne, pour chaque item j, la probabilité d'observer la modalité de réponse $x = 0, 1, ..., M_j - 1$ conditionnellement aux paramètres de l'item et au niveau de variable latente de l'individu i. Le modèle s'écrit :

$$P(X_{ij} = x \mid \theta_i, \delta_{j1}, \delta_{j2}, ..., \delta_{jM_j-1}) = \frac{\exp(x\theta_i - \sum_{p=1}^{x} \delta_{jp})}{\sum_{l=0}^{M_j-1} \exp(l\theta_i - \sum_{p=1}^{l} \delta_{jp})} \tag{2.15}$$

Avec un PCM, les items ne sont plus caractérisés par un seul paramètre de seuil, mais par plusieurs : les paramètres δ_{jp}. Pour un item j donné, on dénombre $M_j - 1$ paramètres de seuil (autant que de modalités de réponse strictement supérieures à 0). Soit $p \in \{1, ..., M_j - 1\}$, le paramètre de seuil δ_{jp} correspond au niveau de variable latente pour lequel les modalités de réponse adjacentes p et $p-1$ sont équiprobables. Plus un paramètre de seuil δ_{jp} est élevé, plus la modalité de réponse p est considérée comme difficile à atteindre.

Pour un item j donné, on peut représenter graphiquement la probabilité d'observer chacune des modalités de réponse en fonction du niveau de la variable latente. Ces courbes sont appelées courbes caractéristiques des modalités de réponse (*Category Characteristic Curves*, CCC). Un exemple est proposé dans la figure 2.10 avec un item polytomique ayant quatre modalités de réponse. Les paramètres de seuil de l'item correspondent aux niveaux de variable latente pour

lesquels les courbes de deux catégories de réponse adjacentes s'intersectent. Un diagramme à barres empilées est parfois inséré sous ces CCC pour indiquer la modalité de réponse la plus probable pour chaque niveau de variable latente. Ce n'est pas le cas pour la figure 2.10, mais cette pratique sera parfois employée dans le manuscrit.

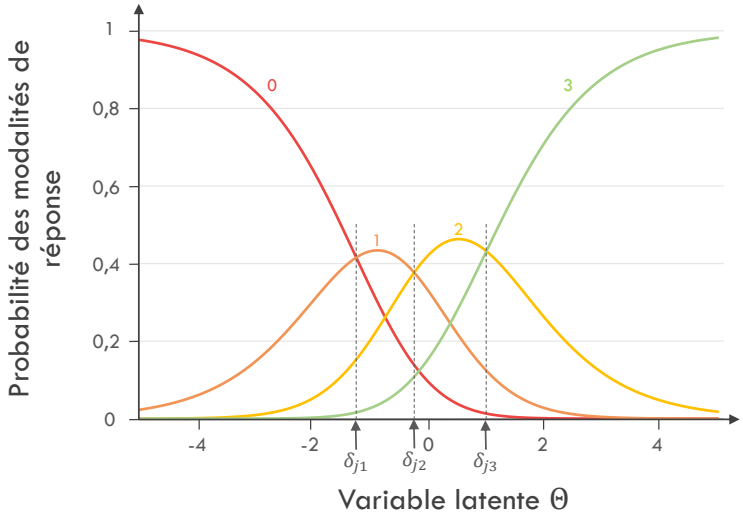


FIGURE 2.10 – Courbes caractéristiques des modalités de réponse d'un item j polytomique (modalités : 0, 1, 2 et 3) selon le modèle de crédit partiel (PCM)

On peut également calculer l'espérance du score de l'item pour chaque niveau de variable latente. Elle est définie par :

$$E(X_{ij} \mid \theta_i) = \sum_{p=0}^{M_j-1} p \times P(X_{ij} = p \mid \theta_i) \tag{2.16}$$

Dans la littérature, cette espérance est désignée sous le terme d'*expected item score function* (fonction du score attendu à l'item) ou encore de *true-score function*. Sa représentation graphique est souvent appelée courbe caractéristique de l'item (comme pour les items dichotomiques). La courbe caractéristique de l'item considéré dans la figure 2.10 est donnée en figure 2.11. On peut remarquer que :

– Plus le niveau de variable latente de l'individu i est élevé, plus l'espérance du score à l'item est grande.

– Pour des niveaux de variable latente très faibles (vers $-\infty$), l'espérance du score à l'item tend vers 0 (asymptote horizontale à gauche).

– Pour des niveaux de variable latente très grands (vers $+\infty$), l'espérance du score à l'item tend vers $M_j - 1$ le score maximal atteignable (asymptote horizontale à droite).

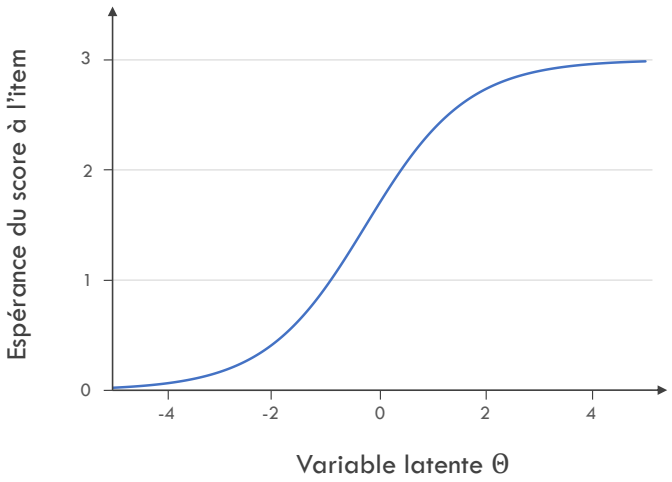


FIGURE 2.11 – Courbe caractéristique d'un item polytomique selon le modèle de crédit partiel (PCM)

Estimation des modèles de la famille de Rasch

L'estimation des paramètres d'un modèle de Rasch ou d'un PCM dépend de la façon dont est considérée la variable latente. Elle peut soit être vue comme un paramètre fixe ou bien comme un effet aléatoire. Lorsque la variable latente est considérée comme un effet aléatoire, on suppose généralement qu'elle est normalement distribuée. De plus, pour que le modèle soit identifiable, une possibilité est de supposer que la moyenne de cette distribution est égale à 0 (contrainte d'identifiabilité [3]). Les paramètres peuvent ensuite être estimés par maximum de vraisemblance marginale (*Marginal Maximum Likelihood*, MML) [52]. C'est le choix qui a été fait dans tous les travaux du manuscrit impliquant des modèles de la famille de Rasch.

3. Il s'agit là d'une contrainte d'identifiabilité possible. Il en existe une autre qui porte sur la moyenne des paramètres de seuil, supposée être égale à zéro.

2.3.3 Les modèles de la famille de Lord

Contemporainement aux travaux de Rasch, le psychométricien Frederic Lord a proposé une vision de la psychométrie plus centrée sur la caractérisation des items. Le paradigme est différent de celui de Rasch puisque l'objectif est ici de proposer un modèle qui s'adapte au mieux aux items (et non pas rechercher des items qui satisfont un modèle précis). Par conséquent, les modèles qui découlent de la vision de Lord ne bénéficient pas des propriétés des modèles issus de la théorie de la mesure de Rasch. Deux modèles sont présentés dans cette section : le 2-PLM et le modèle de crédit partiel généralisé (*Generalized Partial Credit Model*, GPCM). Il en existe cependant beaucoup d'autres.

Le modèle logistique à deux paramètres pour items dichotomiques

Le modèle logistique à deux paramètres (*Two-Parameter Logistic Model*, 2-PLM) a été proposé par Birnbaum [41]. Ce modèle est conçu pour des items dichotomiques, et permet de considérer que la "pente" de la courbe caractéristique de l'item j puisse différer de celle de l'item j'.

La formulation mathématique du 2-PLM est proche de celle du modèle de Rasch, mais les items sont cette fois-ci caractérisés par deux paramètres :

- Le paramètre de seuil de l'item δ_j (*item threshold / location / difficulty*)
- Le paramètre de discrimination de l'item α_j (*discrimination parameter*)

Le modèle 2-PLM s'écrit :

$$P(X_{ij} = 1 \mid \theta_i, \alpha_j, \delta_j) = \frac{\exp(\alpha_j(\theta_i - \delta_j))}{1 + \exp(\alpha_j(\theta_i - \delta_j))} \tag{2.17}$$

Comme pour le modèle de Rasch, le paramètre de seuil correspond au niveau de variable latente pour lequel la probabilité de répondre 1 à l'item vaut 50%. Ce qui distingue le 2-PLM du modèle de Rasch, c'est que le 2-PLM permet de considérer que certains items sont plus discriminants que d'autres (grâce à l'introduction d'un paramètre de discrimination α_j propre à chaque item). Ce paramètre de discrimination influe sur la pente de la courbe caractéristique de l'item au point d'inflexion. Plus le paramètre de discrimination est grand, plus la pente sera

marquée et plus l'item sera discriminant (c.-à-d., qu'il permettra de différencier des individus ayant des niveaux de variable latente différents). Avec ce modèle, les courbes caractéristiques de différents items peuvent être sécantes (ce n'était pas le cas avec le modèle de Rasch). À titre d'exemple, plusieurs courbes caractéristiques d'items ayant différents paramètres de discrimination sont données dans la figure 2.12 (le paramètre de seuil de ces items est maintenu constant pour faciliter l'interprétation de ces courbes).

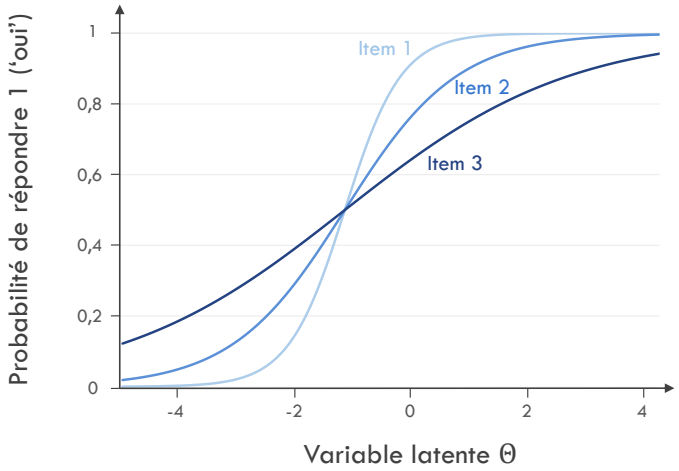


FIGURE 2.12 – Courbes caractéristiques de trois items dichotomiques vérifiant le modèle logistique à deux paramètres (2-PLM)

Notes : l'item 1 est plus discriminant que l'item 2, qui est lui-même plus discriminant que l'item 3.

Le modèle de crédit partiel généralisé pour items polytomiques

Le modèle de crédit partiel généralisé pour items polytomiques (*Generalized Partial Credit Model*, GPCM) a été proposé par Muraki [53]. Sa formulation mathématique ressemble à celle d'un PCM, avec l'ajout d'un paramètre de discrimination α_j pour chaque item.

Considérons un ensemble de J items polytomiques ayant chacun M_j modalités de réponse. Le GPCM donne, pour chaque item j, la probabilité d'observer la modalité de réponse $x = 0, 1, ..., M_j - 1$ conditionnellement aux paramètres de l'item et au niveau de variable latente de l'individu i.

Le modèle s'écrit :

$$P(X_{ij} = x \mid \theta_i, \alpha_j, \delta_{j1}, \delta_{j2}, ..., \delta_{jM_j-1}) = \frac{\exp(\alpha_j(x\theta_i - \sum_{p=1}^{x} \delta_{jp}))}{\sum_{l=0}^{M_j-1} \exp(\alpha_j(l\theta_i - \sum_{p=1}^{l} \delta_j p))} \tag{2.18}$$

Comme pour le PCM, on peut tracer les courbes caractéristiques de chacune des modalités de réponse d'un item. Les paramètres de seuil d'un item correspondent ici encore aux niveaux de variable latente pour lesquels les courbes de deux modalités de réponse adjacentes s'intersectent. Enfin, il est également possible de tracer les courbes caractéristiques des items en calculant l'espérance du score aux items pour chaque niveau de variable latente.

Estimation des modèles de la famille de Lord

Comme les modèles de la famille de Rasch, les modèles de la famille de Lord peuvent être estimés par maximum de vraisemblance marginale.

2.3.4 Hypothèses fondamentales des modèles de l'IRT et de la RMT

Les modèles de l'IRT et de la RMT présentés précédemment reposent sur trois hypothèses fondamentales qui sont : (i) l'unidimensionnalité, (ii) la monotonicité et (iii) l'indépendance locale.

- L'hypothèse d'**unidimensionnalit**é spécifie que l'ensemble des items considérés est censé mesurer un seul et même construit psychologique. Des extensions ont été proposées pour modéliser des questionnaires multidimensionnels, mais il n'en sera pas question dans ce manuscrit.
- La **monotonicité** correspond au fait que l'espérance du score à un item donné doit augmenter lorsque le niveau de la variable latente augmente.
- L'**indépendance locale** signifie que les réponses des individus aux items sont indépendantes, conditionnellement à la variable latente. Si les réponses des individus à un item sont influencées par leurs réponses à d'autres items, on parle de dépendance locale.

Chapitre 3

Présentation du DIF et du *Response shift*

Sommaire

3.1 Problématique de la non-invariance de la mesure

Lorsque l'on analyse des données rapportées par les patients, on cherche souvent à comparer les niveaux du concept mesuré entre des groupes et/ou au cours du temps. Pour pouvoir réaliser des comparaisons valides, il faut néanmoins s'assurer que l'hypothèse d'invariance de la mesure entre groupes et au cours du temps est respectée. Prenons pour exemple un questionnaire conçu pour mesurer la dépression et soumis à un même groupe d'individus à deux temps de mesure. Pour que l'hypothèse d'invariance de la mesure soit respectée, il faut que :

- Les individus ayant des niveaux de dépression très proches répondent de façon similaire aux items du questionnaire, et ce, indépendamment de leurs caractéristiques (comme l'âge, le sexe, le contexte culturel, l'ethnie, etc.). On parle d'invariance entre groupes d'individus.
- Les réponses d'un même individu entre les deux temps de mesure reflètent bien l'évolution réelle du niveau de dépression de l'individu : c'est-à-dire des réponses similaires si le niveau de dépression est stable et des réponses qui s'améliorent (ou se détériorent) si le niveau de dépression diminue (ou augmente) au cours du temps. On parle alors d'invariance longitudinale.

En termes plus techniques, l'hypothèse d'invariance de la mesure entre groupes et au cours du temps spécifie que les réponses aux items d'un questionnaire dépendent uniquement du niveau de la variable latente ciblée par ce questionnaire [15]. Ainsi, si une association est trouvée entre les réponses aux items et une variable de groupe, cette association doit pouvoir s'expliquer par une différence de niveau de variable latente entre les groupes. De même, si une association est trouvée entre les réponses aux items et le temps, cette association doit pouvoir s'expliquer par une évolution du niveau de la variable latente.

En pratique, il peut néanmoins y avoir des différences dans la façon dont les individus perçoivent et interprètent le construit ciblé par le questionnaire (et les items associés) en raison de leurs caractéristiques culturelles, environnementales et personnelles, mais également des expériences qu'ils ont vécues. En reprenant l'exemple de questionnaires mesurant la dépression, plusieurs études réalisées au début des années 2000 ont par exemple suggéré qu'à niveau de dé-

pression égal, les hommes tendent à sous-déclarer leurs pleurs par rapport aux femmes (voir par exemple les références [54–56]). On parle alors de fonctionnement différentiel des items (*Differential Item Functioning*, DIF).

De plus, lors d'un suivi longitudinal, la perception et l'interprétation de certains items peuvent également changer au cours du temps. Par exemple, les stratégies thérapeutiques de l'état dépressif visent souvent à modifier le processus de référence à soi pour rectifier le biais cognitif négatif du patient (comme la prédominance des contenus négatifs dans les pensées de l'individu, la sur-généralisation des éléments négatifs, la maximisation du négatif et la minimisation du positif, etc.) [57]. Ces stratégies peuvent donc perturber la mesure longitudinale de la dépression en modifiant les normes de mesure internes de certains patients et leur façon de répondre aux questionnaires [6]. Si de tels changements de normes internes ont lieu entre deux temps de mesure, les réponses obtenues à chaque temps ne sont alors pas directement comparables : le changement observé ne reflétant pas uniquement le changement que l'on souhaitait mesurer au niveau du construit. Cette dissonance est connue dans la littérature sous le nom de *response shift*.

Le *response shift* et le DIF sont deux cas particuliers de non-invariance de la mesure. Pour le *response shift*, l'association entre les réponses aux items et la variable latente dépend du temps, alors que pour le DIF, cette association dépend d'une covariable de groupe (le sexe pour l'exemple des pleurs).

En présence de DIF et/ou de *response shift*, les données recueillies ne sont pas directement comparables entre les groupes ou dans le temps. Ignorer ces phénomènes peut entraîner un biais de mesure dans les comparaisons réalisées et donc : une mésinterprétation des données recueillies, des conclusions erronées et des prises de décision cliniques potentiellement inadéquates [16]. À terme, les attentes et les besoins des individus vis-à-vis des soins pourraient ne pas être satisfaits [16, 58]. En reprenant l'exemple de la dépression, la comparaison brute des scores de dépression observés entre les hommes et les femmes pourrait notamment indiquer à tort que les femmes sont plus déprimées que les hommes (cette conclusion erronée étant due à une sous-déclaration des pleurs par les hommes par rapport aux femmes, à niveau de dépression égal par ailleurs). De façon similaire, lors d'un suivi longitudinal, la comparaison brute des scores obtenus à deux

temps de mesure pourrait à la fois représenter un changement du niveau du construit mesuré, mais également un changement dans la perception et l'interprétation du questionnaire.

Le DIF et le *response shift* sont également des critères d'intérêt en eux-mêmes. Pour le DIF, il peut être intéressant de chercher à comprendre pourquoi certains items fonctionnent différemment chez certains individus [59] :

- L'item mesure-t-il un autre concept en plus de celui auquel on s'intéresse ?
- Quelle est la covariable réellement à l'origine de ce fonctionnement différentiel (DIF causal) [20, 60].
- En cas de comparaisons interculturelles : en quoi les groupes comparés sont-ils différents ? Y a-t-il eu un problème de traduction et/ou d'adaptation transculturelle du questionnaire ?

Quant au *response shift*, ce phénomène est considéré dans la littérature comme une résultante de la façon dont les patients font face et s'adapte aux changements liés à leurs conditions de santé [18, 61] et mérite donc d'être étudié en tant que tel. Il est donc capital de comprendre pourquoi ce phénomène survient chez certains individus afin de pouvoir élaborer des leviers d'action pour aider les patients à faire face aux événements qu'ils rencontrent et mieux vivre avec leurs nouvelles conditions de vie (et ainsi éviter les situations de maladaptation). Sawatzky *et al.* [58] et Mayo [62] ont par ailleurs rappelé que le *response shift* pourrait être un objectif de soins en soi, par exemple en soins de réhabilitation ou encore en thérapies cognitives et comportementales .

Dans la suite de ce chapitre, les concepts de DIF et de *response shift* sont présentés plus en détails, avec un aperçu général non exhaustif des méthodes qui permettent de s'y intéresser.

3.2 Le fonctionnement différentiel des items

3.2.1 Présentation du concept

Le DIF survient lorsqu'un ou plusieurs items d'un questionnaire "fonctionnent" différemment entre les groupes que l'on cherche à comparer. Plus précisément, cela signifie que ces items ne reflètent pas de façon équivalente le construit étudié entre les groupes [16]. Dans ce cas, la distribution des réponses aux items ne dépend pas uniquement du construit que l'on souhaite mesurer, mais également des caractéristiques des répondants [15] (comme le sexe pour l'item sur les pleurs utilisé pour mesurer la dépression). Le fait que certains items fonctionnent différemment entre des groupes d'individus peut s'expliquer par des différences d'attentes, d'interprétation, de culture ou de personnalité [16].

Le terme "fonctionnement différentiel des items" (*differential item functioning*) est apparu pour la première fois en 1986 dans les travaux de Holland et Thayer [63]. Néanmoins, l'idée que certains items fonctionnent différemment entre des groupes d'individus est plus ancienne. Elle a été introduite dans le cadre de la théorie de réponse à l'item sous le terme d'*item bias* [64], et on la retrouve notamment dans les travaux de Mellenbergh [15, 65] et de Lord [64]. Dans ces travaux, le DIF (ou *item bias*) est défini de façon très générale, mais il est en fait illustré dans le cas simple d'un item dichotomique fonctionnant différemment entre deux groupes d'individus. De façon formelle, un item dichotomique X (visant à mesurer la variable latente Θ) fonctionne différemment entre deux groupes définis par une covariable G si la probabilité de "réussir l'item" (c.-à-d. répondre 1) diffère entre les deux groupes, à niveau de variable latente égal par ailleurs. Mathématiquement, cela s'écrit :

$$P(X = 1|G = 0, \Theta = \theta) \neq P(X = 1|G = 1, \Theta = \theta) \tag{3.1}$$

Autrement dit, les paramètres de la fonction reliant le niveau de variable latente des individus à leurs réponses au questionnaire diffèrent entre les groupes considérés. Cela se traduit graphiquement par des courbes caractéristiques d'items qui ne coïncident pas (voir figure 3.1 pour le cas d'un item dichotomique avec des courbes caractéristiques de type logistique).

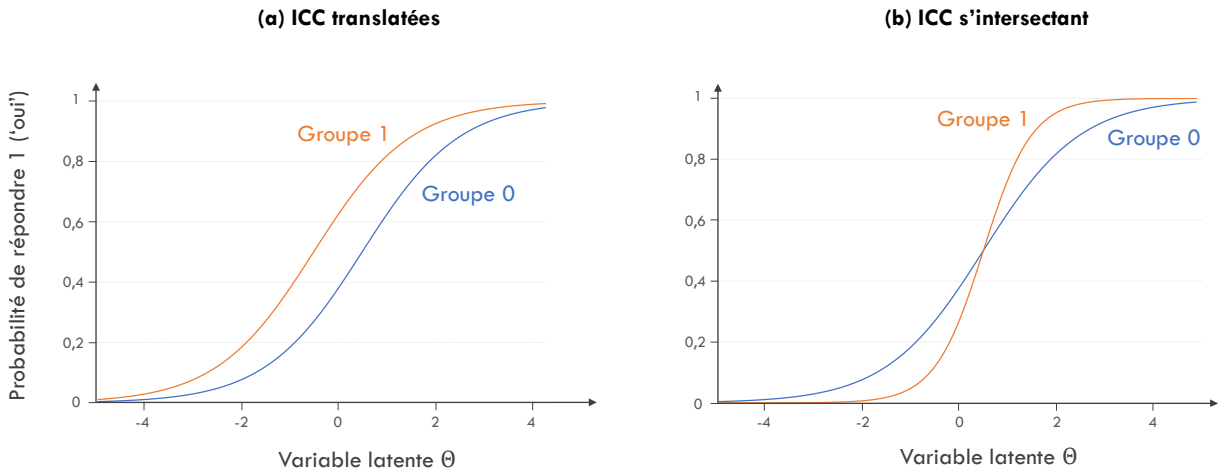


FIGURE 3.1 – Exemple de courbes caractéristiques d'items (*Item Caracteristic Curve*, ICC) ne coïncidant pas entre deux groupes d'individus (DIF)

Deux formes de DIF ont été décrites par Mellenbergh [66] :

- Le **DIF uniforme :** la relation statistique entre la variable de groupe et la réponse à l'item est la même quel que soit le niveau de la variable latente ;
- Le **DIF non uniforme :** la relation statistique entre la variable de groupe et la réponse à l'item varie en fonction du niveau de la variable latente.

Néanmoins, comme l'a rapporté Hanson, la définition du DIF uniforme n'est pas consensuelle dans la littérature [67]. Il en découle nécessairement une instabilité de la définition du DIF non uniforme. Certains auteurs ont préféré la terminologie de DIF "unidirectionnel ou bidirectionnel" [68–70] :

- Le **DIF unidirectionnel** est une forme de DIF où l'écart entre les ICC au sein des groupes étudiés conserve le même signe quel que soit le niveau de variable latente. Lorsqu'on étudie le DIF entre deux groupes d'individus, cette forme de DIF correspond au cas où une courbe caractéristique est toujours au-dessus de l'autre. Le DIF unidirectionnel n'est pas forcément uniforme (la relation entre la variable de groupe et la réponse à l'item pouvant changer, tant qu'il n'y a pas d'inversion dans la relation d'ordre entre les courbes). En revanche, le DIF uniforme tel que défini par Mellenbergh est forcément unidirectionnel [70].

– Le **DIF bidirectionnel** est une forme de DIF où l'écart entre les ICC au sein des groupes étudiés change de signe.

Un exemple pour ces formes de DIF est donné dans la figure 3.1. Le graphique (a) présente deux courbes caractéristiques d'un même item translatées entre les deux groupes : le DIF est alors unidirectionnel. Le graphique (b) présente deux courbes caractéristiques d'un même item qui s'intersectent et pour lesquelles la relation d'ordre s'inverse.

En pratique, avec le modèle de Rasch pour items dichotomiques, le DIF est forcément uniforme. En effet, la courbe caractéristique d'un item j n'est définie que par un seul paramètre : le paramètre de seuil δ_j. Ce modèle ne permet donc de considérer que le cas du DIF où les courbes caractéristiques des items sont translatées (figure 3.1, graphique (a)).
Pour le 2-PLM, la courbe caractéristique d'un item j est cette fois définie par deux paramètres : le paramètre de seuil δ_j et le paramètre de discrimination α_j. Une différence entre groupes dans le paramètre de seuil entraînera une translation des courbes caractéristiques, tandis qu'une différence entre groupes au niveau du paramètre de discrimination se manifestera par une différence de pente au niveau du point d'inflexion des courbes caractéristiques (courbes qui, par conséquent, se croiseront : voir figure 3.1, graphique (b)). Ces deux formes de DIF peuvent survenir simultanément. Dans la littérature, une différence au niveau du paramètre de seuil δ_j est généralement qualifiée de DIF uniforme, tandis qu'une différence au niveau du paramètre de discrimination α_j est qualifiée de DIF non uniforme (le terme DIF non uniforme est en fait plus large et englobe également le cas où les deux paramètres δ_j et α_j diffèrent tous deux entre les groupes étudiés).

Si les formes de DIF sont relativement limitées avec des items dichotomiques, elles sont beaucoup plus nombreuses dans le cas d'items polytomiques. En effet, le DIF peut se manifester sur toutes les modalités de réponse, ou uniquement certaines. Il peut changer de magnitude en fonction de la modalité de réponse touchée, et même changer de sens. En termes d'opérationnalisation, le DIF sur un item polytomique j peut toucher les paramètres associés à ses modalités de réponse (c.-à-d. les paramètres de seuil δ_{jp} pour les modèles de la famille de Rasch et ceux de la famille de Lord). Le DIF peut également interférer avec le paramètre de discrimination de l'item (pour les modèles de la famille de Lord uniquement).

La figure 3.2 illustre quelques formes de DIF possibles sur les paramètres de seuil d'un item à $M = 4$ modalités entre deux groupes d'individus (PCM) :

1. Graphique (a) : les paramètres de seuil de l'item dans le groupe 1 sont tous décalés vers la droite par rapport à ceux du groupe 0. La magnitude du décalage est la même.
2. Graphique (b) : les paramètres de seuil de l'item dans le groupe 1 sont tous décalés vers la droite par rapport à ceux du groupe 0. La magnitude du décalage augmente.
3. Graphique (c) : les paramètres de seuil de l'item dans le groupe 1 sont décalés par rapport à ceux du groupe 0, le sens du décalage varie en fonction du paramètre d'item considéré. La magnitude du décalage varie également.

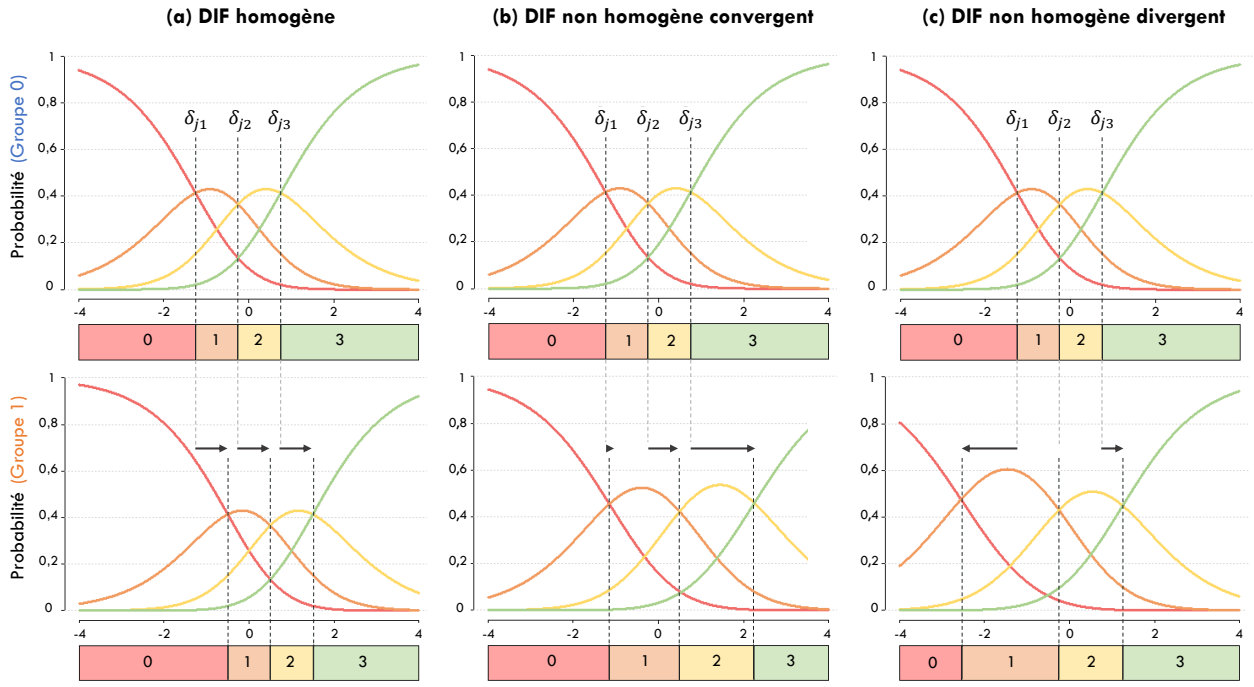


FIGURE 3.2 – Formes possibles de DIF (item polytomique à 4 modalités, *Partial Credit Model*)
Notes :
– *Les flèches noires représentent les différences de paramètres de seuil entre groupes*
– *L'axe des abscisses représente le niveau de la variable latente* Θ

Dans le manuscrit, les formes décrites dans les points (2) et (3) seront appelées DIF non homogène [71]. On peut noter que ces deux formes de DIF illustrent respectivement le *differential step functioning* (DSF) convergent et divergent de la taxonomie proposée par Penfield *et al.* [72]. Pour maintenir une terminologie cohérente tout au long du manuscrit, la forme de DIF décrite dans le point (1) sera appelée DIF homogène (DSF systématique et convergent pour Penfield *et al.* [72]).

Si au niveau des courbes caractéristiques des modalités, les possibilités pour la forme du DIF sont multiples, on retrouve néanmoins la notion de DIF unidirectionnel ou bidirectionnel en traçant les courbes caractéristiques des items. Les courbes caractéristiques des items associées aux exemples considérés dans la figure 3.2 sont représentées ci-dessous (figure 3.3). On peut remarquer que lorsque le DIF est homogène, la courbe caractéristique de l'item dans le groupe 1 est translatée par rapport à celle du groupe 0 (courbes parallèles, graphique (a)). Pour le DIF non homogène convergent, les deux courbes caractéristiques de l'item ne se croisent pas (DIF unidirectionnel : la courbe bleue [groupe 0] est toujours au-dessus de la courbe orange [groupe 1], graphique (b)). Enfin, dans notre exemple, lorsque le DIF est non homogène divergent, les deux courbes caractéristiques de l'item se croisent (DIF bidirectionnel, graphique (c)).

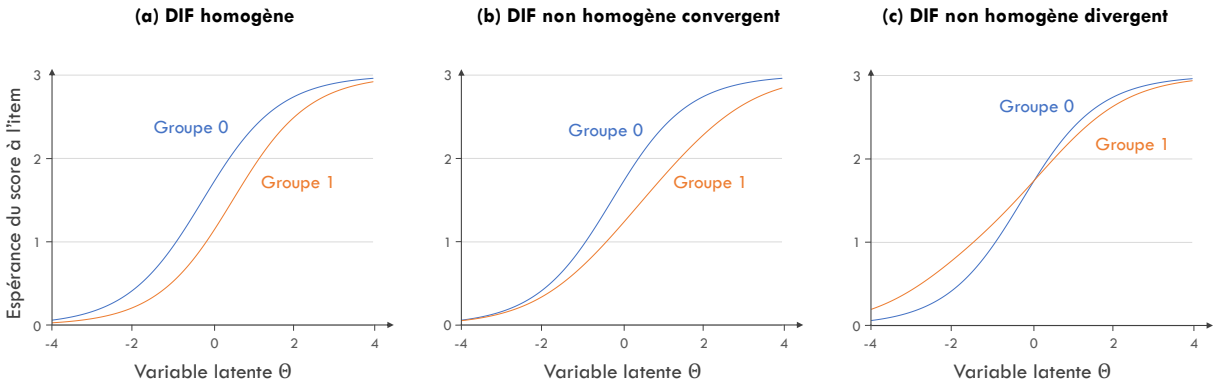


FIGURE 3.3 – Courbe caractéristique d'un item polytomique à 4 modalités affecté par différentes formes de DIF (*Partial Credit Model)*

La recherche de DIF est un point primordial à prendre en compte lors du développement et de la traduction d'un questionnaire. Il en va de même lors de l'analyse de données issues de questionnaire. En effet, la présence de DIF peut rendre la comparaison des groupes vis-à-vis du construit étudié ambiguë : est-ce que la différence entre groupes observée est "réelle", ou s'agit-il d'un "artéfact" (comme une interprétation différente des items par les membres des différents groupes étudiés) [73] ?

Aussi, la détection du DIF a été une question très étudiée au cours de ces 20/30 dernières années et de nombreuses méthodes ont été proposées. Parmi les plus connues, on peut notamment citer deux méthodes "historiques" : la méthode de Mantel-Haenszel [74] et la procédure basée sur la régression logistique [75]. Pour les résumer succinctement, ces deux méthodes cherchent à détecter la présence de DIF en regroupant les individus en fonction de leur score total (utilisé comme un *proxy* de leur niveau vis-à-vis du construit étudié). Il s'agit ensuite de déterminer si, à score total égal par ailleurs, il existe une différence (entre les groupes considérés) dans la probabilité de réussir un item donné. L'utilisation du score total comme variable de "*matching*" a depuis été remise en question (notamment par Millsap et Everson [76]). L'une des raisons invoquées est que le score total risque d'être biaisé si une large proportion d'items fonctionnent différemment entre les deux groupes [76].

Dans le champ de la santé (mais également de façon plus générale), l'utilisation de modèles à variable latente pour la détection du DIF est actuellement plébiscitée [21]. Plutôt que d'énumérer l'ensemble des méthodes de détection du DIF qui ont pu être proposées par le passé (ce qui a déjà été fait dans la littérature [70, 77–79]), la section suivante du manuscrit présente les méthodes de détection du DIF dans le cadre de la théorie de réponse à l'item (IRT) et la théorie de la mesure de Rasch (RMT), qui permettent de prendre en compte simultanément différentes sources de DIF possibles. Autrement dit, nous nous intéresserons ici aux méthodes qui permettent de considérer en même temps plusieurs covariables lors de la recherche du DIF (et qui sont basées sur l'estimation de modèles issus de la famille de Rasch ou de Lord). Ces méthodes sont d'intérêt, car lorsque l'on analyse des données autorapportées, les variables que l'on suspecte être à l'origine d'un fonctionnement différentiel peuvent être multiples et il est intéressant de tenter de démêler leurs effets pour mieux comprendre ce phénomène [20, 21].

3.2.2 Méthodes de détection statistiques du DIF permettant de considérer plusieurs covariables simultanément (IRT et RMT)

Lorsque l'on cherche à comprendre l'origine du fonctionnement différentiel de certains items, les méthodes de détection du DIF sont assez limitées. En effet, la plupart d'entre elles sont conçues pour considérer une seule covariable à la fois [19, 20]. Afin d'identifier correctement l'origine du DIF et d'ajuster sur d'éventuels facteurs de confusion, il est nécessaire de pouvoir incorporer simultanément plusieurs covariables dans l'analyse. Les paragraphes suivants dressent un état des lieux des méthodes statistiques de détection du DIF permettant de considérer plusieurs covariables simultanément et étant basées sur l'estimation de modèles de l'IRT ou de la RMT.

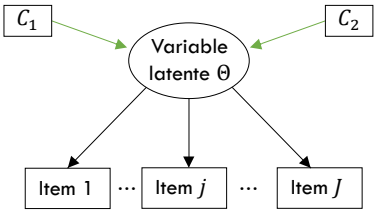
Procédure IRT-C

La procédure IRT-C (*Item Response Theory with Covariates*) a été initialement proposée en 2011 par Tay *et al.* [80] pour détecter du DIF (uniforme ou non uniforme) dans un questionnaire composé d'items dichotomiques. Cette méthode permet de considérer différentes covariables qualitatives et/ou quantitatives simultanément. Elle repose sur l'estimation successive de 2-PLM (modèles de l'IRT), où la détection du DIF est basée sur l'analyse des résidus bivariés (*Bivariate Residuals*, BVR) entre les items et les covariables introduites dans l'analyse. Ces indices seront présentés plus en détail en page 57. Les étapes de cette méthode itérative sont décrites ci-dessous avec une illustration graphique des modèles :

Étape 1 : Modèle initial $M = M_0$

Estimation d'un 2-PLM complètement invariant (supposant l'absence de DIF) où seul l'effet des covariables sur le niveau moyen de la variable latente est estimé.

Illustration avec deux covariables C_1 et C_2



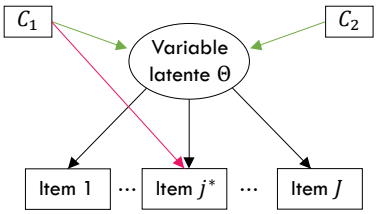
$$P(X_{ij} = 1 | \theta_i, \alpha_j, \delta_j, \beta_1, \beta_2, C_{1_i}, C_{2_i}) = \frac{\exp\left(\alpha_j(\theta_i + \beta_1 C_{1_i} + \beta_2 C_{2_i} - \delta_j)\right)}{1 + \exp\left(\alpha_j(\theta_i + \beta_1 C_{1_i} + \beta_2 C_{2_i} - \delta_j)\right)}$$

Étape 2 : Inspection des résidus bivariés (BVR)

Inspection de la matrice de résidus bivariés associée au modèle M et identification de la paire item-covariable associée à la valeur la plus élevée. Cette paire est notée (item j^*, covariable C^*).

Étape 3 : Test de la présence de DIF uniforme sur la paire identifiée

Estimation d'un nouveau 2-PLM (modèle M-U) à partir du modèle actuel M, où la covariable C^* induit du DIF uniforme sur l'item j^*. Pour ce faire, un nouveau coefficient interférant avec le paramètre de seuil de l'item j^* est introduit dans le modèle. Ce coefficient permet de considérer que le paramètre de seuil de l'item varie entre les différents groupes définis par la covariable C^* (ce coefficient est représenté ci-dessous par le chemin rose dans le cas où $C^* = C_1$). Un test statistique est réalisé pour déterminer si l'effet introduit est significatif (test de Wald).

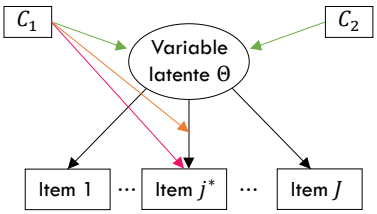


Pour l'item j^* :

$$P\left(X_{ij^*}=1\middle|\theta_i,\delta_{j^*},\alpha_{j^*},\beta_1,\beta_2,C_{1_i},C_{2_i},\gamma_{j^*}^{(C_1)}\right)=\frac{\exp\left(\alpha_j\left(\theta_i+\beta_1C_{1_i}+\beta_2C_{2_i}-[\delta_j+\gamma_{j^*}^{(C_1)}C_{1_i}]\right)\right)}{1+\exp\left(\alpha_j\left(\theta_i+\beta_1C_{1_i}+\beta_2C_{2_i}-[\delta_j+\gamma_{j^*}^{(C_1)}C_{1_i}]\right)\right)}$$

Étape 4 : Test de la présence de DIF non uniforme sur la paire identifiée

Estimation d'un nouveau 2-PLM (modèle M-NU) à partir du modèle M-U, où la covariable C^* induit du DIF uniforme et non uniforme sur l'item j^*. Pour ce faire, un nouveau coefficient interférant avec le paramètre de discrimination de l'item j^* est introduit dans le modèle. Ce coefficient permet de considérer que le paramètre de discrimination de l'item varie entre les différents groupes définis par la covariable C^* (ce coefficient est représenté ci-dessous par le chemin orange dans le cas où $C^* = C_1$). Un test statistique est réalisé pour déterminer si l'effet introduit est significatif (test de Wald).



Pour l'item j^* :

$$P\left(X_{ij^*}=1\middle|\theta_i,\delta_{j^*},\alpha_{j^*},\beta_1,\beta_2,C_{1_i},C_{2_i},\gamma_{j^*}^{(C_1)},\lambda_{j^*}^{(C_1)}\right)=\frac{\exp\left(\left(\alpha_j+\lambda_j^{(C_1)}C_{1_i}\right)\left(\theta_i+\beta_1C_{1_i}+\beta_2C_{2_i}-[\delta_j+\gamma_{j^*}^{(C_1)}C_{1_i}]\right)\right)}{1+\exp\left(\left(\alpha_j+\lambda_j^{(C_1)}C_{1_i}\right)\left(\theta_i+\beta_1C_{1_i}+\beta_2C_{2_i}-[\delta_j+\gamma_{j^*}^{(C_1)}C_{1_i}]\right)\right)}$$

Étape 5 : Mise à jour du modèle

Le modèle M est mis à jour en fonction des résultats trouvés au cours des deux étapes précédentes. Dans ce modèle, seuls les effets DIF associés à un test de Wald significatif sont conservés. Néanmoins, il convient de noter que si l'effet modélisant une différence entre groupes au niveau du paramètre de discrimination est conservé, alors le coefficient "associé" qui modélise le DIF uniforme est, lui aussi, conservé (que son test soit significatif ou non).

Itérations : La procédure est répétée à partir de l'étape 2 en se basant sur ce nouveau modèle et en investiguant les autres paires item-covariable. Elle s'arrête lorsque les valeurs de BVR obtenues sont suffisamment faibles (sans qu'un seuil ne soit spécifié par les auteurs). Bien que ce ne soit pas indiqué, la procédure doit également probablement s'arrêter avant de modéliser du DIF sur tous les items pour une covariable donnée (sinon l'effet de la covariable sur le niveau moyen de la variable latente ne peut plus être estimé, problème d'identifiabilité).

Fin de la procédure : Parmi tous les modèles M estimés au cours de la procédure, le modèle retenu au final est celui présentant le critère d'information bayésien (*Bayesian information criterion*, BIC) [81] et le critère d'information d'Akaike consistent (*Consistent Akaike Information Criterion*, CAIC) [82] les plus faibles. Les effets DIF détectés correspondent aux effets modélisés dans le modèle M conservé.

Modifications de la procédure : Cette méthode a par la suite fait l'objet de deux autres articles, où ses étapes ont été légèrement remaniées :

– En 2013, les étapes de la procédure ont été reformulées et légèrement remaniées par Tay *et al.* [20]. Premièrement, les étapes 3 (DIF uniforme) et 4 (DIF non uniforme) ont été permutées, et l'étape "DIF uniforme" était omise si du DIF non uniforme était mis en évidence. De plus, l'étape "critère d'arrêt avec le BIC" a été clairement formalisée et ajoutée avant l'étape de mise à jour du modèle. Dans cette étape "critère d'arrêt", les BIC des modèles *M-U* et *M-NU* sont comparés à celui du modèle actuel M. Si les deux modèles intermédiaires présentent des BIC plus élevés, alors la procédure s'arrête et le modèle M est sélectionné comme modèle final. Sinon le modèle est mis à jour en fonction des formes de DIF mises en évidence. Les étapes sont ensuite répétées jusqu'à ne plus mettre en évidence de DIF ou jusqu'à observer une augmentation du BIC. Une alternative utilisant l'AIC3 [83] au lieu du BIC a également été proposée. Tay *et al.* indiquent que cette étape "critère d'arrêt" avec le BIC ou l'AIC3 est facultative : la procédure peut aussi continuer tant que du DIF est mis en évidence. Par ailleurs, comme précédemment, la procédure s'arrête probablement avant de modéliser du DIF sur tous les items pour une covariable donnée (sinon, on rencontre un problème d'identifiabilité du modèle, mais ce n'est pas indiqué par les auteurs).

– En 2016, cette procédure a été étudiée par Tay *et al.* [84] dans le cadre du modèle logistique à trois paramètres (*Three-Parameter Logistic Model*, 3-PLM). Il s'agit d'un modèle peu utilisé en santé, mais fréquemment recontré dans les sciences de l'éducation. Comme le 2-PLM, ce modèle est issu de l'IRT et est conçu pour des items binaires. Il présente un paramètre d'item en plus par rapport au 2-PLM : le paramètre de *pseudo-guessing* c_j. Ce paramètre permet de considérer que la probabilité de réussir un item j tend vers une valeur c_j potentiellement non nulle quand la variable latente est infiniment petite (tends vers $-\infty$). Dans cet article, la procédure a été adaptée pour ne rechercher que du DIF uniforme (différence entre groupes dans les paramètres de seuil de certains items).

Si les performances de cette procédure ont été évaluée à plusieurs reprises par simulations dans le cadre de questionnaire composé d'items dichotomiques, elle n'a en revanche jamais été étendue aux items polytomiques.

Résidus bivariés : Les résidus bivariés, sur lesquels sont basés cette procédure, sont présentés par Tay *et al.* comme étant "analogues aux indices de modification des modèles à équations structurelles SEM" [20, 80]. Ces indices de modification sont souvent plébiscités pour étudier l'invariance de la mesure (voir par exemple les travaux de B.O Muthén [85, 86] et de Oort [87, 88]). Une valeur élevée de résidus bivariés entre un item et une covariable indiquerait une dépendance locale entre l'item et la covariable, ce qui pourrait signifier la présence de DIF [84]. Ces résidus bivariés sont des mesures ressemblant à un chi-deux de Pearson et reflétant l'écart entre les effectifs attendus et observés dans le tableau de contingence croisant une covariable de groupe et un item. Ils quantifient l'association covariable-item qui n'est pas expliquée par le modèle [20]. En 2015, Tay *et al.* [89] ont indiqué que cette méthode serait analogue à la méthode MIMIC (voir paragraphe suivant). Néanmoins, l'utilisation des BVR pour identifier des hypothèses d'invariance potentiellement problématiques a été remise en question par Oberski *et al.* [90]. En effet, la distribution asymptotique des résidus bivariés n'est pas connue, et des valeurs faibles ne doivent pas être considérées comme indicatives d'un bon ajustement. Par ailleurs, ces indices sont peu implémentés dans les logiciels statistiques (à la suite d'un modèle de l'IRT ou de la RMT nous ne sommes parvenus à les obtenir qu'avec le logiciel *Latent Gold*). Cela complexifie donc l'implémentation de la méthode et limite son utilisation.

Gestion des variables qualitatives à trois modalités ou plus : Ce type de variable semble être introduit dans l'analyse à l'aide de variables indicatrices binaires créées après avoir : (i) choisi une modalité de référence et (ii) réalisé un codage disjonctif complet pour les modalités restantes.

Méthode MIMIC

L'utilisation des modèles MIMIC (*Multiple Indicators Multiple Causes*) pour la détection du DIF a été initialement proposée par B.O. Muthén dans les années 1980 [86, 91, 92]. L'intérêt autour des modèles MIMIC a néanmoins connu un véritable essor dans les années 2010, notamment grâce aux travaux de Woods [93–96]. Les modèles MIMIC sont souvent reliés aux modèles à équations structurelles SEM. Néanmoins, les modèles MIMIC peuvent également être paramétrisés en modèles de l'IRT ou de la RMT : Woods a par exemple utilisé la méthode MIMIC pour

détecter du DIF à partir d'un 2-PLM [93, 96]. Cette section restera focalisée sur les modèles IRT/RMT, mais la distinction entre les SEM et ces modèles n'est parfois plus réellement faite par certains auteurs [22].

Les modèles MIMIC sont caractérisés par l'inclusion de covariables observées (quantitatives ou qualitatives [1]) dans le modèle étudié. Ces covariables peuvent à la fois influencer le niveau moyen de la variable latente et les paramètres des items qui lui sont associés, grâce à l'introduction de nouveaux coefficients dans le modèle (comme pour la procédure IRT-C, cette dernière s'étant inspirée de la méthode MIMIC). Du DIF est mis en évidence si une covariable influence significativement le paramètre de discrimination et/ou les paramètres de seuil d'un item.

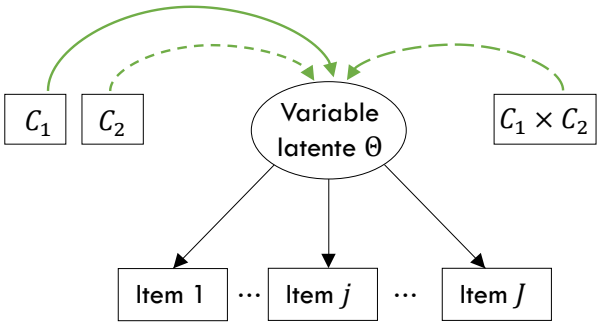
La méthode MIMIC est aujourd'hui présentée dans la littérature comme une méthode très flexible qui permet de rechercher du DIF en prenant en compte plusieurs covariables simultanément (et leurs éventuelles interaction) [97–99]. Si cette assertion est techniquement vraie, il faut néanmoins la nuancer pour plusieurs raisons.
Tout d'abord, il n'y a pas vraiment de consensus sur la façon d'utiliser les modèles MIMIC pour détecter le DIF [93] (il y a en fait de nombreuses façons de faire parfois très différentes qui coexistent et sont désignées sous le même nom).
De plus, peu d'études se sont intéressées aux performances des méthodes type "MIMIC" pour la recherche de DIF en présence de plusieurs covariables. Chun *et al.* [22] ont d'ailleurs dressé le constat suivant en 2016 : la capacité des modèles MIMIC à détecter du DIF en présence de plusieurs covariables et à identifier les covariables à l'origine du DIF n'a jamais été évaluée par simulations [22]. Aussi, ces auteurs ont proposé (et évalué par simulations) trois implémentations de la méthode MIMIC pour détecter du DIF en présence de deux covariables binaires et leur interaction. Ces implémentations sont respectivement nommées *Constrained baseline*, *Free baseline* et *New sequential-free baseline*. Nous allons nous concentrer sur cette dernière implémentation dans les paragraphes suivants.

1. Les variables à trois modalités ou plus sont introduites sous la forme de variables indicatrices binaires.

L'implémentation de la méthode MIMIC "*New sequential-free baseline*" pour détecter du DIF en présence de deux covariables binaires C_1 et C_2 se déroule en trois étapes, présentées ci-dessous. Pour se placer dans un cadre IRT, elles ont été ici adaptées pour considérer un GPCM (cela n'entraîne pas de perte de généralité). Les modèles ne sont pas écrits, car peu lisibles.

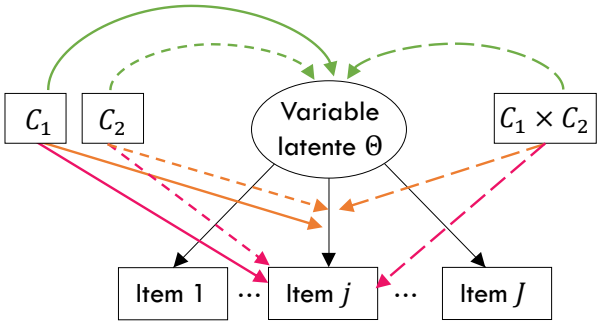
Étape n°1 : Modèle de référence M_0

La première étape consiste à construire un GPCM sur lequel sera basé l'ensemble des étapes suivantes. Dans ce modèle, deux covariables binaires C_1 et C_2 (ainsi que leur interaction $C_1 \times C_2$) sont intégrées de façon à impacter le niveau moyen de la variable latente. Ces effets sont représentés par les flèches vertes.



Étape n°2 : Identification d'un item sans DIF (item dit *anchor*)

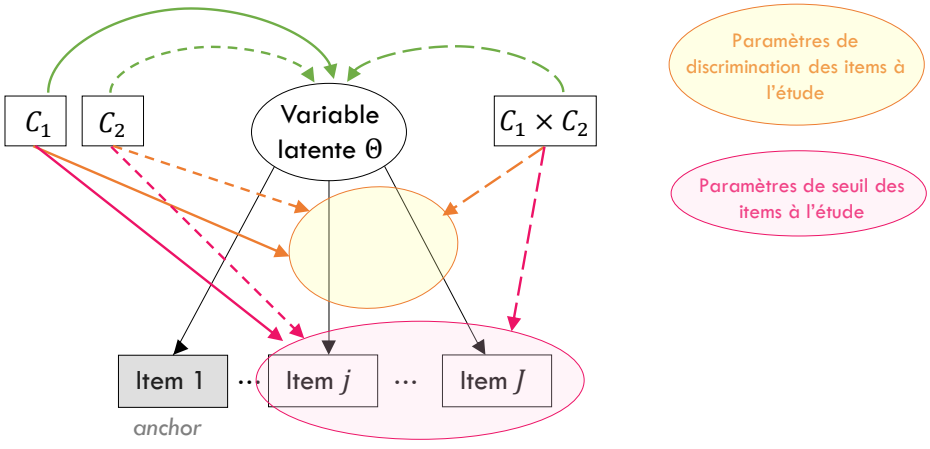
L'objectif est d'identifier l'item le plus discriminant qui n'est pas affecté par du DIF. Pour ce faire, un GPCM est estimé pour chaque item j (modèle M_j) où tous les items sont invariants, sauf l'item j étudié. En termes MIMIC, cela signifie que les effets de C_1, C_2 et $C_1 \times C_2$ sur les paramètres de l'item j sont tous estimés. L'effet de ces variables sur les paramètres de seuil de l'item sont représentés en rose, et les effets sur le paramètre de discrimination sont représentés en orange. La méthode suppose que l'effet de chaque covariable sur les paramètres de seuil δ_{jp} de l'item est constant pour tout $p \geq 1$ (DIF homogène). L'intégration de ces effets permet de modéliser du DIF sur l'item j induit par les deux covariables et leur interaction. Aucun autre paramètre modélisant du DIF n'est intégré.



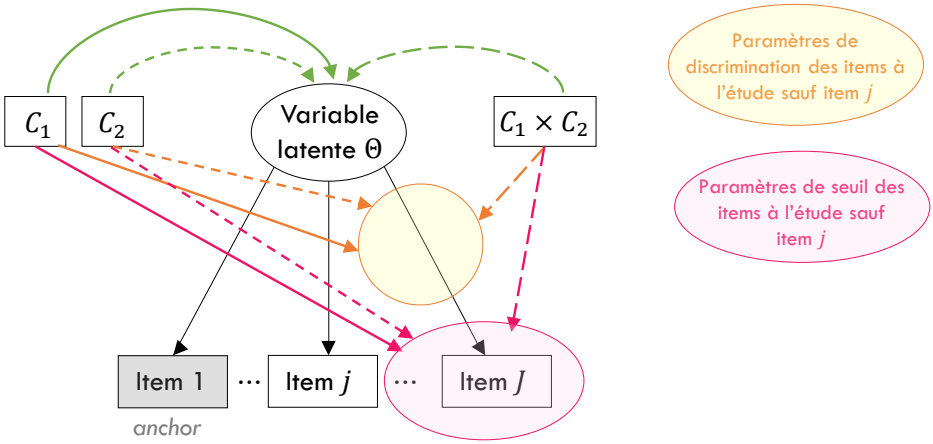
Un test du rapport de vraisemblance comparant les modèles emboîtés M_0 et M_j est réalisé pour déterminer le modèle ayant la meilleure adéquation. Si le test est non significatif (risque de première espèce fixé à 5%), l'item est alors candidat pour être l'item *anchor* de l'étape suivante. Une fois tous les modèles estimés (et les items testés), il faut identifier, parmi les items candidats pour être *anchor*, celui qui a le paramètre de discrimination le plus élevé. Par simplicité, on considèrera qu'il s'agit de l'item 1 pour la suite de la procédure. Cet item servira d'item *anchor* pour l'étape suivante. Les autres items seront désignés sous le terme d'"items à l'étude".

Étape n°3 : Identification des items affectés par du DIF

L'objectif est d'identifier les items affectés par du DIF parmi les items à l'étude. Pour ce faire, un premier modèle (modèle $\widetilde{M}_0$) est estimé de façon à ce que seul l'item identifié comme anchor (item 1) soit invariant. En termes MIMIC, cela signifie que les effets de C_1, C_2 et $C_1 \times C_2$ sur les paramètres de tous les items à l'étude sont estimés. L'intégration de ces effets permet de modéliser du DIF induit par les deux covariables et leur interaction sur tous les items à l'étude.



L'étape 3 se poursuit par l'estimation de plusieurs GPCM (un pour chaque item j à l'étude : Modèle $\widetilde{M}_j$). Ces modèles sont basés sur le modèle $\widetilde{M}_0$, mais l'effet des covariables sur les paramètres de l'item j n'est cette fois-ci plus estimé (on suppose l'absence de DIF sur l'item j). Chaque modèle $\widetilde{M}_j$ est comparé au modèle $\widetilde{M}_0$ par un test du rapport de vraisemblance. On conclut qu'un item est affecté par du DIF si ce test est significatif à 5%.



Autres implémentations : Les deux autres implémentations sont assez proches de celle présentée. L'implémentation *constrained baseline* correspond au fait de ne faire que les étapes 1 et 2 (les items affectés par du DIF sont ceux ayant un test significatif lors de l'étape 2). L'implémentation *free baseline* correspond au fait de ne faire que les étapes 1 et 3 (l'item *anchor* étant supposé connu *a priori* par la personne menant l'analyse).

Étude de l'origine du DIF : Chun *et al.* [22] indiquent qu'ils ont cherché à déterminer si les implémentations de la méthode MIMIC qu'ils proposent permettent de retrouver correctement l'origine du DIF. Pour ce faire, ils ont réalisé des tests du rapport de vraisemblances pour chaque item j à l'étude. Les modèles comparés étaient :

- Modèle $M_j^{(C_1,C_2,C_1\times C_2)}$: modèle qui contient tous les effets des covariables (C_1, C_2, $C_1 \times C_2$) sur les seuils et le paramètre de discrimination de l'item j ;
- Modèle $M_j^{(C_1,C_2)}$: modèle qui contient uniquement les effets des covariables C_1 et C_2 sur les paramètres de l'item j (effets de l'interaction $C_1 \times C_2$ supprimés) ;

– Modèle $M_j^{(C_2)}$: modèle qui contient uniquement les effets de la covariable C_2 sur les paramètres de l'item j (effets de C_1 supprimé) ;

– Modèle $M_j^{(\varnothing)}$: modèle où aucune des covariables n'a d'effet sur les paramètres de l'item j.

Chun *et al.* concluait que du DIF était induit par l'interaction sur l'item j si le test du rapport de vraisemblance comparant $M_j^{(C_1,C_2,C_1\times C_2)}$ à $M_j^{(C_1,C_2)}$ était significatif.
Du DIF était induit par la covariable C_1 si le test comparant $M_j^{(C_1,C_2)}$ à $M_j^{(C_2)}$ était significatif.
Idem pour la covariable C_2 en comparant $M_j^{(C_2)}$ à $M_j^{(\varnothing)}$.
Néanmoins, ces auteurs n'indiquent pas à quoi correspond le modèle initial $M_j^{(C_1,C_2,C_1\times C_2)}$:
si l'on sait que les effets des covariables sur les paramètres de seuil et de discrimination de l'item à l'étude sont impactés par les covariables, on ne sait pas ce qu'il en est pour les autres items. De plus, l'étude de simulation menée indique que les implémentations considérées présentaient de faibles performances pour retrouver correctement la source du DIF.

Forme du DIF : Les implémentations proposées par Chun *et al.* [22] ne permettent pas de considérer que l'effet d'une covariable sur les paramètres de seuil d'un item varie en fonction du seuil considéré. Un seul effet est modélisé pour tous les paramètres de seuil d'un item donné (DIF homogène). Ces méthodes ne permettent donc pas d'investiguer le DIF non homogène sur les paramètres de seuil.

Remarque sur les items *anchor* : Lorsque l'on recherche du DIF, il n'est pas possible d'estimer l'effet des covariables sur le niveau moyen de la variable latente tout en supposant que ces covariables induisent du DIF sur tous les items (le modèle n'est pas identifiable). C'est pourquoi les méthodes de recherche de DIF s'appuient sur la désignation, ou la détermination, d'items *anchor* (items que l'on supposera fonctionner de la même façon pour tout le monde). Ces items peuvent soit être désignés *a priori* (grâce à des connaissances accumulées sur le questionnaire ou sur le contexte clinique) ou déterminés à l'aide de tests statistiques comme c'est le cas pour l'implémentation *New sequential-free baseline* des MIMIC présentée ci-dessus.

GPCM-Lasso et PCM-Lasso

Pour pouvoir considérer simultanément plusieurs covariables lors de la détection du DIF, Schauberger et Mair ont proposé une méthode basée sur l'estimation pénalisée de modèles à variable latente (PCM et GPCM) [19]. Cette méthode permet de considérer des items polytomiques et fait suite aux travaux de Tutz et Schauberger sur items binaires (modèle de Rasch) [100]. Sur la même idée que la méthode MIMIC, le DIF est modélisé par l'introduction de coefficients interférant avec les paramètres des items. Schauberger et Mair ne se sont néanmoins intéressés qu'au DIF sur les paramètres de seuil. Deux méthodes ont en fait été proposées par ces auteurs. Elles sont présentées ci-dessous dans le cas d'un GPCM. On considèrera ici encore deux covariables C_1 et C_2 (binaires ou continues [2]).

– La première méthode suppose que le DIF est homogène, aussi l'effet d'une covariable sur les différents paramètres de seuil d'un item polytomique est constant.

  Le modèle à estimer s'écrit alors pour tout item j :

$$P\left(X_{ij} = x \mid \theta_i, \delta_{j1}, \ldots, \delta_{jM_j-1}, \alpha_j, C_{1_i}, C_{2_i}, \beta_1, \beta_2, \gamma_j^{(C_1)}, \gamma_j^{(C_2)}\right) = \frac{\exp\left(\alpha_j\left(x\left[\theta_i + \beta_1 C_{1_i} + \beta_2 C_{2_i}\right] - \sum_{p=1}^{x}\left[\delta_{jp} + \gamma_j^{(C_1)} C_{1_i} + \gamma_j^{(C_2)} C_{2_i}\right]\right)\right)}{\sum_{l=0}^{M_j-1}\exp\left(\alpha_j\left(l\left[\theta_i + \beta_1 C_{1_i} + \beta_2 C_{2_i}\right] - \sum_{p=1}^{l}\left[\delta_{jp} + \gamma_j^{(C_1)} C_{1_i} + \gamma_j^{(C_2)} C_{2_i}\right]\right)\right)} \tag{3.2}$$

– La seconde ne fait pas cette hypothèse. Dans ce cas, le modèle à estimer s'écrit :

$$P\left(X_{ij} = x \mid \theta_i, \delta_{j1}, \ldots, \delta_{jM_j-1}, \alpha_j, C_{1_i}, C_{2_i}, \beta_1, \beta_2, \gamma_{j1}^{(C_1)}, \ldots, \gamma_{jM_j-1}^{(C_1)}, \gamma_{j1}^{(C_2)}, \ldots, \gamma_{jM_j-1}^{(C_2)}\right) = \frac{\exp\left(\alpha_j\left(x\left[\theta_i + \beta_1 C_{1_i} + \beta_2 C_{2_i}\right] - \sum_{p=1}^{x}\left[\delta_{jp} + \gamma_{jp}^{(C_1)} C_{1_i} + \gamma_{jp}^{(C_2)} C_{2_i}\right]\right)\right)}{\sum_{l=0}^{M_j-1}\exp\left(\alpha_j\left(l\left[\theta_i + \beta_1 C_{1_i} + \beta_2 C_{2_i}\right] - \sum_{p=1}^{l}\left[\delta_{jp} + \gamma_{jp}^{(C_1)} C_{1_i} + \gamma_{jp}^{(C_2)} C_{2_i}\right]\right)\right)} \tag{3.3}$$

2. Les variables qualitatives à trois modalités ou plus sont introduites sous la forme de variables indicatrices binaires.

Pour ces deux méthodes, l'objectif est d'estimer les modèles décrits en imposant un "coût" à l'estimation non nulle d'un paramètre de DIF ($\gamma_{jp}^{(C)}$, avec $C = C_1$ ou C_2). Ainsi, pour qu'un paramètre de DIF vaille la peine d'être estimé, le coût engendré par son estimation devra être contrebalancé par un gain de vraisemblance. Sinon, il sera contraint à être nul. On parle d'estimation pénalisée.

C'est grâce à cette pénalisation que les modèles sont estimables. En effet, comme on peut le voir dans les équations des modèles 3.2 et 3.3, du DIF est modélisé pour toutes les paires item-covariable. Ces modèles ne devraient donc pas être identifiables. En imposant une pénalisation sur les paramètres de DIF, Schauberger et Mair [19] s'assurent qu'un nombre suffisant de paramètres de DIF ne seront en vérité pas estimés (contraints à être nuls par la pénalisation), rendant les modèles identifiables. Une fois le modèle estimé, le DIF mis en évidence est indiqué par les paramètres de DIF estimés non nuls. Aucune des deux méthodes ne permet de chercher des interactions entre les effets DIF des covariables.

Seule la méthode supposant le DIF homogène a été évaluée par simulations par Schauberger et Mair [19]. Aucune information sur les performances de l'autre méthode n'est disponible. Ces deux méthodes seront explicitées plus en détail dans le chapitre 5.

DIF item-focused tree

Les DIF *item-focused trees* (DIF-IFT) ont été initialement proposés et évalués par simulation par Tutz et Berger pour détecter du DIF au sein d'items dichotomiques en prenant en compte plusieurs covariables [101]. Cette méthode a ensuite été étendue et évaluée par Bollmann *et al.* dans le cas d'items polytomiques [71]. Cette méthode a été proposée en RMT (elle repose sur l'estimation de modèles de Rasch quand les items sont dichotomiques et sur l'estimation de PCM quand les items sont polytomiques). Il s'agit d'une méthode de détection qui peut être qualifiée d'"itérative hiérarchique". Le principe de cette méthode et ses étapes sont décrites dans la figure 3.4 à l'aide d'un exemple dans lequel du DIF est recherché parmi trois items polytomiques, et où deux covariables binaires (le sexe et la nationalité [Nationalité A / Nationalité B]) sont considérées dans l'analyse.

Méthode DIF-IFT (1/3)

Exemple : $J = 3$ items $M = 3$ modalités de réponses (0,1, 2)
Covariable 1 (C_1) : Sexe (Femme/Homme) , Covariable 2 (C_2) : Nationalité [A / B]

Etape 1 : Estimer un PCM basique, sans introduire de covariable et donc sans DIF (Modèle M).

Etape 2 :
Pour chaque item j et chaque covariable C : estimation d'un PCM « candidat » où l'on suppose que l'item j est affecté par du DIF induit par la covariable C (càd : estimation libre des paramètres de seuil de l'item j dans les deux groupes formés par C).

→ Estimations de 2×3 = 6 modèles candidats

Modèle candidat 1 : *Sexe induit du DIF sur item 1*
Modèle candidat 2 : *Sexe induit du DIF sur item 2*
Modèle candidat 3 : *Sexe induit du DIF sur item 3*
Modèle candidat 4 : *Nationalité induit du DIF sur item 1*
Modèle candidat 5 : *Nationalité induit du DIF sur item 2*
Modèle candidat 6 : *Nationalité induit du DIF sur item 3*

Etape 3 :
Sélection du modèle candidat M* ayant la plus petite déviance.
Le couple item-covariable associé à ce modèle est noté (item j^*, covariable C^*).
Test du rapport de vraisemblance avec correction de Bonferroni comparant le modèle M* au modèle M (risque de première espèce : $\alpha = 5\%/\#(\text{Covariable})$).

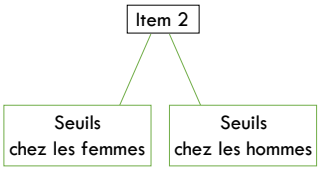
Si test non significatif : L'algorithme s'arrête → pas de DIF mis en évidence : **Modèle final = Modèle initial M.**

Si test significatif :
- DIF détecté (C^* induit du DIF sur l'item j^*) ;
- Mise à jour du modèle M : **M = M*** (les paramètres de seuil de l'item j^* seront définitivement estimés librement entre les deux groupes formés par la covariable C^*).

Pour l'exemple, considérons que le modèle 2 est le modèle ayant la plus petite déviance.
Le test du rapport de vraisemblance permet de déterminer si les paramètres de seuil de l'item 2 diffèrent significativement entre les hommes et les femmes. Ce test est significatif, on détecte alors un DIF sexe sur l'item 2.

Mise à jour du modèle : Modèle M = Modèle 2

On représente graphiquement l'estimation séparée des paramètres de seuil de l'item 2 entre les deux groupes d'individus par la création d'un arbre pour cet item.



Etape 4 : Dérivation de nouveaux modèles candidats à partir du modèle M où un nouvel effet DIF est estimé.

- *Pour les items sur lequel aucun DIF n'a encore été mis en évidence (items j =1, 3)*
  Modèles candidats :
  modèle 1 : Sexe induit du DIF sur l'item 1
  modèle 2 : Sexe induit du DIF sur l'item 3
  modèle 3 : Nationalité induit du DIF sur l'item 1
  modèle 4 : Nationalité induit du DIF sur l'item 3
- *Pour l'item sur lequel du DIF a déjà été mis en évidence (item 2 - DIF induit par le sexe)*
  Modèles candidats :
  modèle 5 : Nationalité induit du DIF sur l'item 2 chez les femmes
  modèle 6 : Nationalité induit du DIF sur l'item 2 chez les hommes

Méthode DIF-IFT (2/3)

Remarque : L'algorithme est hiérarchique, on regarde donc l'effet DIF de la covariable Nationalité sur l'item 2 chez les femmes et chez les hommes séparément.

Etape 5 :
Sélection du modèle candidat M* ayant la plus petite déviance.
Test du rapport de vraisemblance avec correction de Bonferroni comparant le modèle M* au modèle M.

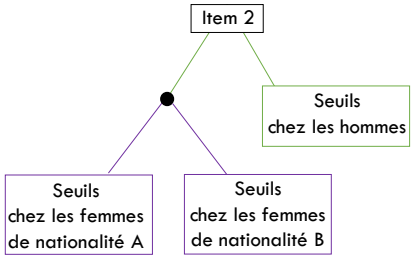
Si test non significatif : L'algorithme s'arrête → **Modèle final = Modèle actuel M.**

Si test significatif :
- Nouvel effet DIF détecté ;
- Mise à jour du modèle M : **M = M*** (la contrainte d'invariance associée à l'effet DIF détecté est définitivement libérée).

Pour l'exemple, considérons que le modèle M ayant la plus petite déviance est celui où la covariable Nationalité induit du DIF sur l'item 2 chez les femmes.*
Le test du rapport de vraisemblance permet de déterminer si les paramètres de seuil de l'item 2 diffèrent significativement entre les femmes de nationalité A et les femmes de nationalité B. Si c'est le cas :

*Mise à jour du modèle : Modèle M = M**

On représente graphiquement l'estimation séparée des paramètres de seuil de l'item 2 entre les femmes de nationalité A et les femmes de nationalité B en ajoutant un niveau à l'arbre déjà créé.



Dans l'exemple, on a considéré qu'on trouvait du DIF sur l'item 2 de façon successive (d'abord pour la covariable sexe, puis pour la covariable nationalité). On aurait pu trouver du DIF sur un autre item à l'étape 5. On aurait alors créé un nouvel arbre pour cet item. Ces arbres permettent d'indiquer, pour chaque item, les sous-groupes pour lesquels différents paramètres de seuil sont estimés. Ces arbres sont néanmoins tous reliés au même modèle qui est mis à jour à chaque étape.

Etapes suivantes : On continue ainsi de suite jusqu'à ne plus trouver de test du rapport de vraisemblance significatif, où jusqu'à avoir atteint une taille de nœud minimale (nombre d'individus dans les sous-groupes mis en évidence qui ne doit pas être inférieur à un nombre défini en amont).

A l'étape suivante les modèles candidats seraient :

- Pour les items sur lequel aucun DIF n'a encore été mis en évidence (items $j = 1, 3$)
  modèle 1 : Sexe induit du DIF sur l'item 1
  modèle 2 : Sexe induit du DIF sur l'item 3
  modèle 3 : Nationalité induit du DIF sur l'item 1
  modèle 4 : Nationalité induit du DIF sur l'item 3

- Pour l'item sur lequel du DIF a déjà été mis en évidence (item 2)
  modèle 6 : Nationalité induit du DIF sur l'item 2 chez les hommes

Méthode DIF-IFT (3/3)

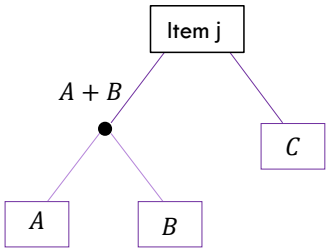
Fonctionnement de l'algorithme avec :

- Plus de deux variables binaires :
Très similaire à ce qui a été présenté mais les arbres pourront être plus grands (ils ne sont pas limités à 4 feuilles comme avec deux covariables binaires).

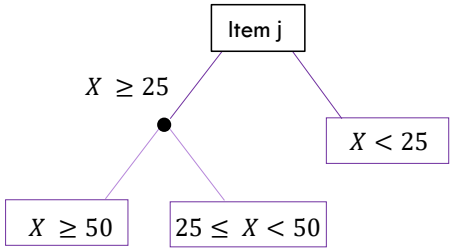
- Avec une variable qualitative à 3 modalités (A, B, C) :
La méthode fait forcément une séparation en deux groupes (arbres binaires), elle recherche donc le meilleur regroupement entre :

- A + B *versus* C
- A + C *versus* B
- B + C *versus* A

En pratique tous les regroupements sont testés (critère de jugement : minimum de déviance).
Contrairement à une covariable binaire, une covariable à trois modalités pourra apparaître à plusieurs reprises dans l'arbre d'un même item :



- Avec une variable ordinale/discrète/continue :
La méthode recherche le seuil s permettant de faire la meilleure séparation en deux groupes :
$X < s$ *versus* $X \geq s$.
En pratique tous les seuils sont testés, critère de jugement : minimum de déviance.
Comme précédemment, la variable pourra revenir à plusieurs reprises dans l'arbre d'un même item, par exemple :



Remarque : Une variable n'est plus candidate pour induire du DIF sur un item donné lorsque tous les seuils sont déjà apparus dans l'arbre de l'item.

FIGURE 3.4 – Méthode DIF *item-focused trees*

Cette méthode permet de considérer que le DIF peut être non homogène. En effet, les différences entre groupes dans les paramètres de seuil des items ne sont pas contraintes à être constantes (les paramètres de seuil des items étant librement estimés entre les groupes). De plus, de par sa philosophie de partitionnement récursif, cette méthode permet de considérer que l'effet DIF d'une covariable varie en fonction d'autres covariables. Par exemple, si l'arbre de l'étape 3 représenté en figure 3.4 est l'arbre final, alors la nationalité induirait du DIF sur l'item 2 uniquement chez les femmes (et pas chez les hommes). DIF-IFT est implémentée sous R (package *DIFtree*), néanmoins, cette méthode ne semble pas permettre l'estimation de l'effet des covariables sur le niveau moyen de la variable latente.

Une synthèse de l'ensemble des méthodes présentées dans cette section est disponible dans le chapitre 5 (figure 5.1).

3.3 Le *response shift*

Les suivis longitudinaux de construits subjectifs comme la qualité de vie liée à la santé, la fatigue, l'anxiété ou la dépression sont aujourd'hui de plus en plus nombreux en recherche clinique, car ils permettent de prendre en compte la perspective et l'expérience des patients. Si l'analyse directe de l'évolution des construits mesurés semble attrayante, elle repose néanmoins sur une hypothèse forte : l'invariance de la mesure longitudinale. Cette hypothèse peut s'avérer invalide, notamment lorsque les individus répondant aux questionnaires sont confrontés à des événements de santé importants au cours de leur suivi. Il est en effet possible que la perception, l'interprétation et la compréhension de certains construits et items changent au cours du temps, à mesure que les patients traversent de nouvelles expériences de vie. Par conséquent, les patients peuvent donner des réponses différentes aux questionnaires entre deux temps de mesure, non seulement parce que leur état de santé a changé, mais également parce que leur référentiel interne a changé. Ces changements de perception dans la façon de s'autoévaluer sont étudiés depuis plus d'une vingtaine d'années dans le champ de la santé et sont connus sous le nom de *response shift*.

3.3.1 Les origines du *response shift*

Historiquement, le terme "*response shift*" est apparu en 1979 dans les travaux de Howard *et al.* [102]. À l'époque, ces auteurs ont publié des résultats qu'ils jugeaient "paradoxaux", obtenus alors qu'ils cherchaient à évaluer expérimentalement l'efficacité d'un atelier de formation aux techniques de communication visant à réduire le dogmatisme. Afin d'évaluer quantitativement cette formation, les individus y participant devaient remplir une échelle de dogmatisme au début de la première session (mesure pré-intervention) puis à la fin de la dernière session (mesure post-intervention). Les participants devaient également remplir un formulaire d'évaluation qualitatif de cette formation. Lors de la comparaison des scores pré- et post- intervention, Howard *et al.* ont constaté que les participants rapportaient une augmentation significative du niveau perçu de dogmatisme suite à la formation. Ces résultats ont été jugés paradoxaux, car ils étaient dissonants avec l'objectif de la formation, mais également avec les retours qualitatifs des participants. En effet, la plupart des commentaires rapportés sur le formulaire d'évaluation de la formation indiquaient plutôt des changements s'apparentant à une diminution du dogmatisme.

Lors d'entretiens menés *a posteriori* pour tenter d'expliquer ces résultats, de nombreux participants ont indiqué qu'ils avaient changé de perception sur leur niveau initial de dogmatisme et que la formation leur avait par exemple fait "ouvrir les yeux". La formation aurait donc ainsi permis une meilleure perception et prise de conscience du niveau de dogmatisme. Ce changement de perception dans la façon de s'autoévaluer (ou changement de normes de mesure interne) a alors été appelée "*response shift*".

Dans le domaine du management cette fois-ci, Golembiewski *et al.* [103] ont proposé en 1976 une typologie de changement intra-individuels qui peuvent survenir au cours du temps lors de la complétion d'auto-questionnaire :

- Le changement *alpha* correspond au changement au niveau du construit mesuré ;
- Le changement *beta* se réfère aux changements dans le référentiel (ou les normes internes) du répondant ;
- Le changement *gamma* est défini comme une reconceptualisation du construit étudié.

Quelques années plus tard, dans le champ de la santé, plusieurs études ont mis en évidence des résultats paradoxaux et contre-intuitifs, qui pourraient être interprétés en termes de *response shift*. En 1991, Breetvelt et Van Dam [104] ont notammant relaté plusieurs études basées sur des auto-questionnaires et ayant conclu que les patients atteints d'un cancer ne rapportaient pas de qualité de vie dégradée par rapport aux individus en bonne santé. Face à ces résultats, ces auteurs ont proposé plusieurs tentatives d'explication, parmi lesquelles le *response shift* (défini comme un changement dans les normes internes des patients induit par un événement de santé extrême). Ils concluaient alors leur article en suggérant de considérer avec prudence les réponses aux auto-questionnaires portant sur la qualité de vie et la détresse psychologique [104].

3.3.2 Le *response shift* en santé : définitions et modèles théoriques

Les travaux de Sprangers et Schwartz

En 1999, Sprangers et Schwartz [17] ont combiné et étendu les travaux de Howard *et al.* [102] et Golembiewski *et al.* [103] et ont défini le *response shift* comme "un changement au cours du temps dans la signification de l'autoévaluation d'un construit", qui résulte de trois causes :

- **La recalibration**, c'est-à-dire un changement dans les normes de mesure internes du répondant via lesquelles il interprète les items et les modalités de réponse.

  Exemple : Un individu a l'habitude de consulter un soignant pour ses douleurs chroniques au dos. Lors d'un rendez-vous, il évalue sa douleur à 7/10 (10 correspondant à la pire douleur imaginable). Quelques jours plus tard, cet individu souffre de douleurs extrêmes lors d'une crise de colique néphrétique. Lors de la visite suivante chez son soignant (après s'être rétabli), l'intensité de ses douleurs chroniques au dos est la même qu'à la visite précédente. Néanmoins, les scores de douleur rapportés par le patient aux deux temps de mesure ne coïncident pas. En effet, il rapporte cette fois-ci un score de 4/10. En fait, les douleurs causées par la crise de colique néphrétique ont été si intenses qu'elles lui ont donné une nouvelle référence pour "la pire douleur imaginable". L'individu a revu ses normes internes : sa façon d'évaluer un même niveau de douleur a changé (cf. figure 3.5) [18].

- **La repriorisation** définie comme un changement dans l'importance relative accordée aux dimensions constituant le construit mesuré.

  Exemple : Un individu a subi un accident de la route duquel il conserve des séquelles physiques permanentes. Avant son accident, cet individu donnait autant d'importance à son fonctionnement physique qu'à ses relations avec les autres. Quelques mois après, même s'il accorde encore de l'importance à son fonctionnement physique, il trouve que le fait de passer des moments privilégiés avec ses proches est au final ce qui lui importe le plus (cf. figure 3.6) [18].

- **La reconceptualisation**, qui correspond à un changement de définition du construit mesuré.

  Exemple : Un individu vient d'apprendre qu'il est atteint d'un cancer avec un mauvais pronostic de survie à court terme. Au début de son suivi, ce patient peut définir sa qualité de vie comme l'absence de symptômes, un bon fonctionnement social et un bon fonctionnement émotionnel. Plus tard, suite à l'initiation de soins de fin de vie, il est possible que la signification de la qualité de vie change pour cet individu, avec l'apparition de nouvelles composantes comme la spiritualité, le sentiment de préparation à la fin de vie et la bonne communication avec le personnel soignant (cf. figure 3.7).

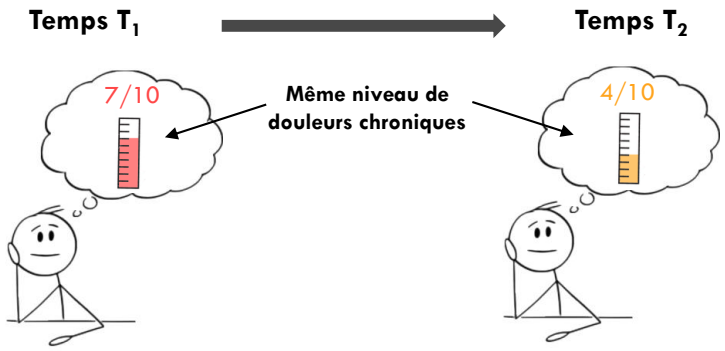


FIGURE 3.5 – Un individu a changé ses normes de mesure internes après avoir connu des épisodes de douleurs intenses (recalibration)

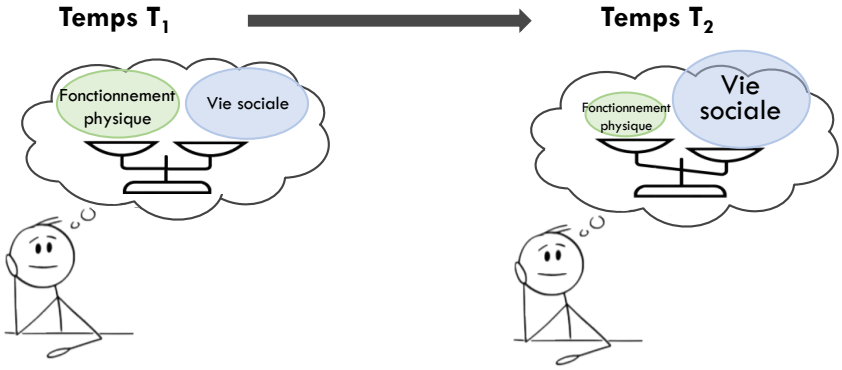


FIGURE 3.6 – Un individu a changé ses valeurs après avoir subi une blessure lui laissant des séquelles physiques permanentes (repriorisation)

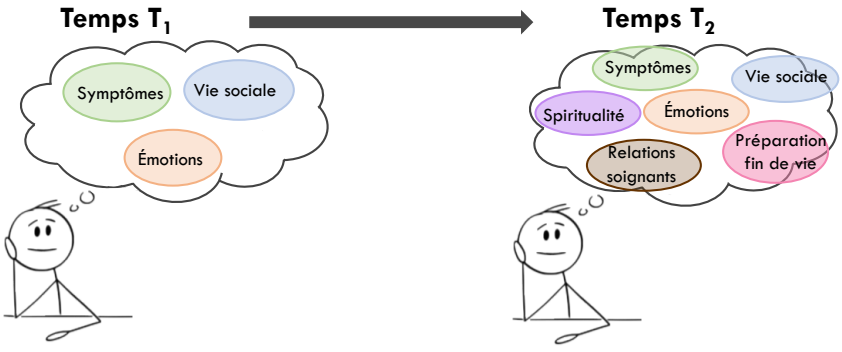


FIGURE 3.7 – Un individu a changé sa définition de la qualité de vie, après avoir appris que son espérance de vie était limitée à quelques mois suite au diagnostic d'un cancer agressif (reconceptualisation)

Dans leur article de 1999, Sprangers et Schwartz ont également proposé un premier modèle théorique pour décrire comment le *response shift* peut affecter l'étude de l'évolution de la qualité de vie liée à la santé suite à un changement d'état de santé [17]. Ce modèle et ses composantes sont présentés en détail dans la figure 3.8.

Dans ce modèle, un catalyseur, correspondant à un changement d'état de santé, peut engendrer chez la personne interrogée des mécanismes psychologiques mis en place afin de s'adapter au changement vécu (comme du *coping* ou de la comparaison sociale). Ces mécanismes psychologiques peuvent ensuite à leur tour engendrer du *response shift* (recalibration, repriorisation et reconceptualisation), qui pourrait affecter l'autoévaluation du construit mesuré (la qualité de vie dans le modèle de Sprangers et Schwartz) [17]. Par ailleurs, les mécanismes adoptés par un individu pour faire face au catalyseur et le *response shift* qui en pourrait en résulter dépendrait des caractéristiques des personnes interrogées (désignées sous le terme d'antécédents). Une boucle de rétroaction indique que ce processus serait itératif et dynamique : des changements dans les normes internes, les valeurs et la définition du construit mesuré pourraient entraîner une réinitialisation des mécanismes établis et l'apparition de nouveaux.

On peut voir dans la figure 3.8 que le *response shift* est considéré comme un concept à part entière qui peut être isolé des mécanismes et du construit mesuré. Le *response shift* pourrait permettre de mieux comprendre les changements de qualité de vie observés au cours du temps [17]. Pour ces auteures, le concept de *response shift* n'a pas pour but de remplacer les théories existantes comme la théorie de l'adaptation [105], la théorie du *coping* [106], la théorie de l'incertitude face à la maladie [107], etc. Il s'agit au contraire d'un concept à part entière pouvant être intégré aux théories existantes [17].

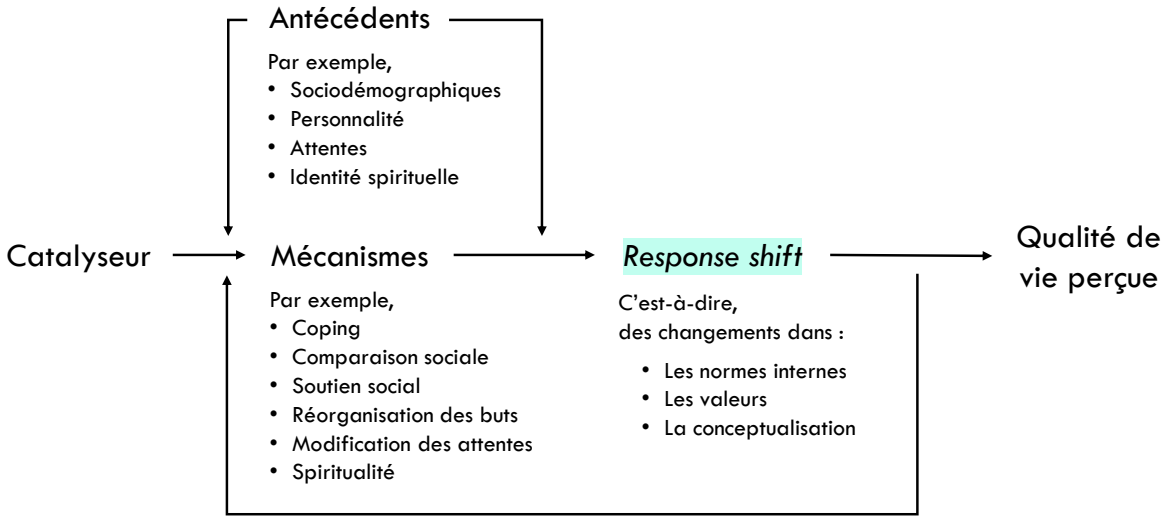


FIGURE 3.8 – Premier modèle théorique du *response shift* et de la qualité de vie proposé par Sprangers et Schwartz
Source : Adapté de Sprangers et Schwartz, Social Sciences and Medicine, 1999

Notes : Ce modèle est construit autour de cinq grandes composantes définies ci-dessous

- *Le **catalyseur** désigne un changement dans l'état de santé de la personne interrogée ;*
- *Les **antécédents** correspondent à des caractéristiques stables de la personne interrogée ;*
- *Les **mécanismes** englobent les processus comportementaux, cognitifs et affectifs qui permettent à la personne interrogée de s'adapter au catalyseur ;*
- *Le **response shift** correspond aux changements (chez la personne interrogée) des normes internes, des valeurs, et de la définition du construit étudié ;*
- *La **qualité de vie perçue** fait référence à un construit multidimensionnel, comprenant au moins trois dimensions principales : le fonctionnement physique, psychologique et social.*

Les travaux de Rapkin et Schwartz

En 2004, Rapkin et Schwartz [108] ont proposé une mise à jour du modèle théorique du *response shift* prenant appui sur les travaux de Tourangeau *et al.* [109]. Le modèle qu'ils proposent est centré sur les processus cognitifs utilisés par les individus pour répondre aux items d'un questionnaire. Ces processus cognitifs sont désignés sous le terme d'*appraisal*.

Rapkin et Schwartz ont décrit quatre grands processus cognitifs à l'œuvre lorsqu'un individu répond à un item. Ils sont présentés ci-dessous sous forme d'étapes [108] : Lorsqu'un individu est confronté à un item, il commence tout d'abord par définir un cadre de référence qui englobe les expériences et les événements qu'il trouve pertinents pour sa réponse au moment de l'évaluation (étape 1). Ce cadre de référence dépend notamment du sens qu'il attache à l'item qui lui est présenté. L'individu sélectionne ensuite, à partir de ce cadre de référence, des expériences spécifiques (étape 2) qu'il compare à ses normes de référence afin de les évaluer (étape 3). Enfin, chacune de ces évaluations est combinée et pondérée pour arriver à la formulation d'une réponse (étape 4).

Pour ces auteurs, le *response shift* est lié à la survenue de changements dans ces processus cognitifs (c.-à-d. des changements d'*appraisal*). Le modèle qu'ils proposent est représenté en détail dans la figure 3.9. On y retrouve les antécédents, le catalyseur et les mécanismes déjà présents dans le modèle de Sprangers et Schwartz [17]. On remarque en revanche l'apparition de l'*appraisal*. De plus, l'évolution de la qualité de vie est séparée en deux : une partie qui est expliquée par les influences standard et une partie non expliquée. Ce modèle postule l'existence de trois familles de liens pouvant expliquer l'évolution d'un score de qualité de vie :

– Les influences "standard" sur la qualité de vie : le catalyseur est supposé influencer directement la qualité de vie (S_1) tout comme les antécédents (S_3). Les antécédents peuvent également impacter indirectement la qualité de vie par l'intermédiaire du catalyseur (S_2). Ces chemins sont de couleur bleue dans la figure.

– Les mécanismes de *coping* (ajustement au stress) : le catalyseur est supposé favoriser ou perturber les mécanismes de *coping* (C_1). Ces mécanismes de *coping* peuvent ensuite réduire l'impact du catalyseur sur l'évolution de la qualité (C_3, effet modérateur ou effet "tampon"). Il est à noter que ces mécanismes dépendent potentiellement des caractéristiques des patients (C_2). Ces chemins sont en orange dans la figure.

– Les processus liés à l'*appraisal* : Les mécanismes de *coping* peuvent entraîner des changements dans l'*appraisal* (A_3). Le catalyseur et les antécédents peuvent également influencer l'*appraisal* (A_1 et A_2, respectivement). Ces chemins sont représentés en rose. Des changements d'*appraisal* pourraient affecter l'évaluation de la qualité de vie et conduire à des changements dans les scores observés qui ne peuvent pas être expliqués par les influences "standard". Cet effet de l'appraisal est alors appelé *response shift*.

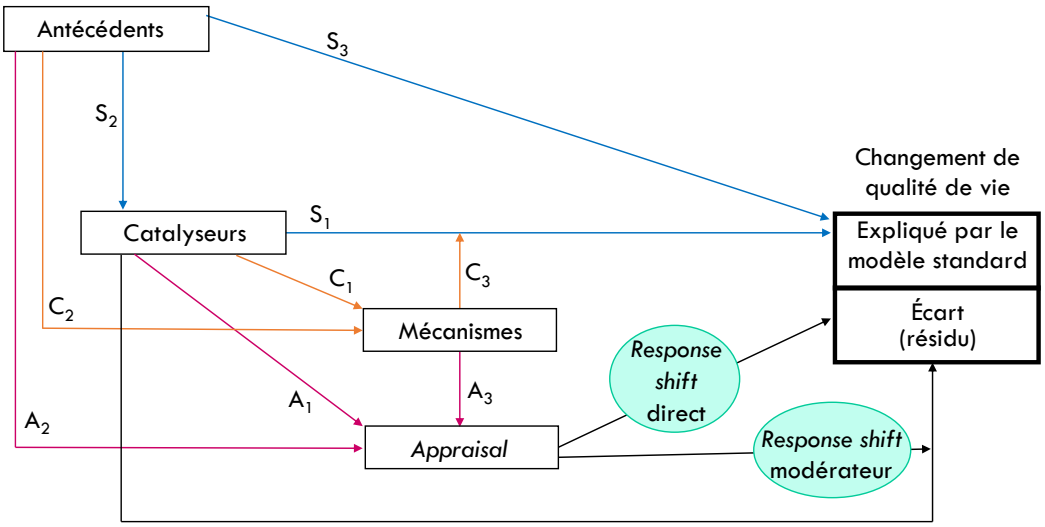


FIGURE 3.9 – Modèle théorique du *response shift* et de la qualité de vie proposé par Rapkin et Schwartz
Source : Adapté de Rapkin et Schwartz, Health and Quality of Life Outcomes, 2004

Dans ce modèle, le *response shift* est donc représenté en termes d'effet déclenché par des changements d'*appraisal* et conduisant à des changements dans les scores observés du construit étudié qui ne peuvent pas être expliqués par les influences "standard". Ces changements d'*appraisal* pourraient être causés par les mécanismes de *coping* ou d'"autres processus" [108].

Pour appuyer leur propos, ces auteurs ont mis en relation les trois types de *response shift* proposés par Sprangers et Schwartz [17] avec les changements d'*appraisal* [108] :

- La reconceptualisation serait liée aux changements dans le cadre de référence (étape 1).
- La repriorisation serait liée aux changements : (i) dans la stratégie de sélection des expériences spécifiques et (ii) dans les facteurs qui déterminent l'importance relative des différentes expériences (étapes 2 et 4).
- La recalibration serait liée aux changements dans les normes de comparaison qui permettent au patient d'évaluer ses expériences (étape 3).

Les travaux de Oort *et al.*

Dans l'objectif de clarifier la définition du *response shift*, Oort et al. ont choisi une perspective assez différente, puisqu'ils ont défini formellement le *response shift* comme un cas spécial de violation du principe d'indépendance conditionnelle [110, 111]. C'est-à-dire une violation du principe selon lequel un questionnaire doit fournir les mêmes résultats à travers différents échantillons ou au cours du temps s'il n'y a pas de différences ou de changements dans le construit ciblé.

Les travaux de Vanier *et al.* (Response shift-In Sync Working group)

Sur la base de l'ensemble des travaux précédemment mentionnés, Vanier *et al.* [18] ont proposé une définition formelle plus spécifique du *response shift*, accompagnée d'un modèle mis à jour. Ces auteurs souhaitaient résoudre certaines difficultés associées aux définitions et modèles préexistants :

– Préciser et clarifier la définition de Oort *et al.* [110, 111] ;

– Faire clairement apparaître les temps de mesure dans le modèle théorique pour clarifier les relations causales entre les composantes du modèle ;

– Distinguer le construit que l'on cherche à mesurer de sa mesure.

Ces auteurs ont choisi de définir le *response shift* comme "un cas spécial de violation du principe de l'indépendance conditionnelle, où le changement de score observé n'est pas entièrement expliqué par l'évolution du construit étudié (le *target change*)". Ce phénomène est supposé être la conséquence d'un "changement dans la signification de l'autoévaluation du construit étudié" (termes issus de la définition du *response shift* de Sprangers et Schwartz [17]).

Vanier *et al.* [18] ont également proposé un modèle théorique décrivant les concepts impliqués dans la survenue du *response shift* et les relations qui les unissent. Ce modèle est représenté dans la figure 3.10 et les chemins reliant les différentes entités sont décrites dans le tableau 3.1. D'après ce modèle, le *response shift* survient lorsque le construit que l'on souhaite mesurer n'explique pas complètement la variabilité des résultats du questionnaire obtenus au deuxième temps de mesure. Deux chemins indiquent la possible survenue de *response shift* :

– L'effet direct du catalyseur sur la réponse au questionnaire au temps 2 (chemin C_3). Par exemple, un choc aigu influençant l'interprétation d'un questionnaire qui serait administré immédiatement après.

– L'effet indirect du catalyseur sur la réponse au questionnaire au temps 2 par l'intermédiaire des mécanismes psychologiques. Cet effet correspond à la combinaison des chemins C_1 et M_2. Il décrit la possibilité que l'adaptation psychologique à une situation impacte la façon dont un individu répond aux items d'un questionnaire au deuxième temps de mesure.

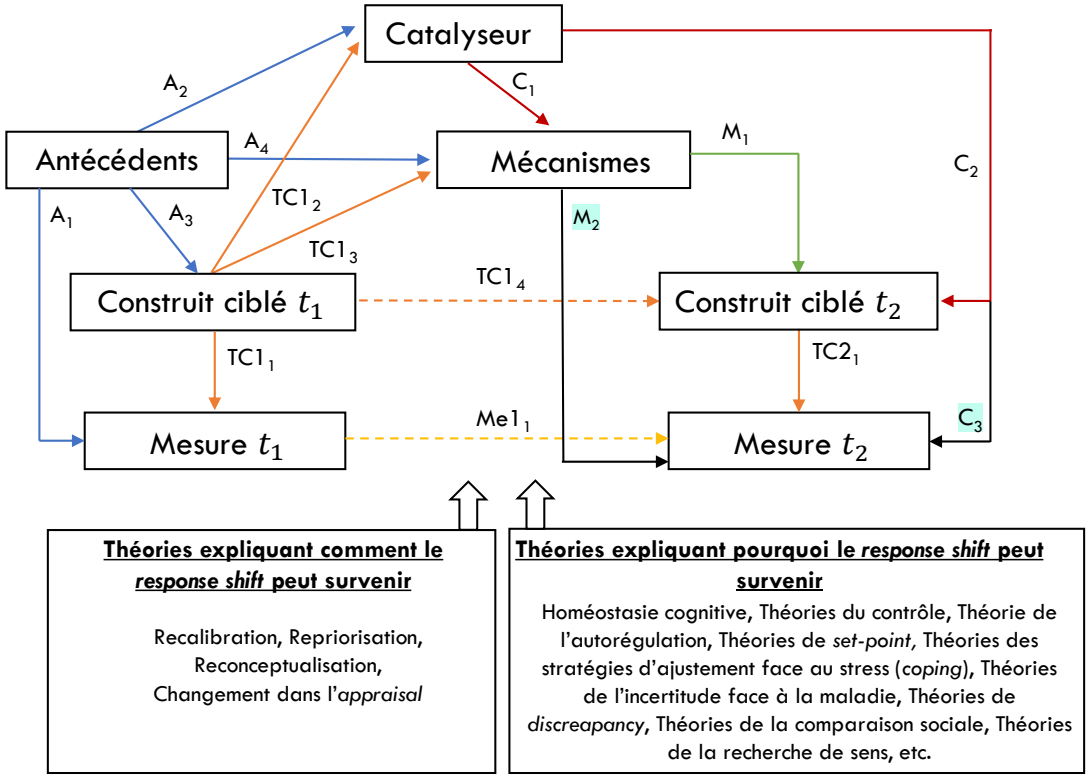


FIGURE 3.10 – Modèle théorique du *response shift* proposé par Vanier *et al.* pour l'analyse de données autorapportées à deux temps de mesure
Source : Adapté de Vanier et al., Quality of Life Research, 2021

Notes :

- t_1 *et* t_2 *: Premier et second temps de mesure, respectivement ;*
- *Le nom et la couleur des flèches indiquent leur origine (par exemple, tous les chemins issus des antécédents ont un nom commençant par A et sont de couleur bleue).*
  Deux exceptions à cette règle : les chemins représentant la potentielle survenue du response shift sont de couleur noire et les chemins issus du construit ciblé sont nommés TC (en référence au nom anglais "Target Construct") ;
- *Catalyseur : Événement de santé ou expérience de vie qui peut impacter le construit étudié au deuxième temps de mesure ;*
- *Antécédents : Caractéristiques plus ou moins stables des individus et de leur environnement ;*
- *Mécanismes : Processus psychologiques (comportementaux, cognitifs et affectifs) que les individus mettent en place pour s'ajuster au catalyseur ;*
- *Construit ciblé : Le construit que l'on cherche à mesurer (comme la qualité de vie) ;*
- *Mesure : Mesure du construit, obtenue à l'aide d'une évaluation autorapportée.*

TABLEAU 3.1 – Description des chemins du modèle proposé par Vanier *et al.*
Source : Adapté de Vanier et al., Quality of Life Research, 2021

Chemins issus des antécédents	
A_1	Les antécédents du répondant peuvent influencer ses réponses à chacun des temps de mesure. (Seul le chemin partant des antécédents vers la mesure au temps 1 est représenté, car l'effet au temps 2 est inclus dans le chemin Me1 1)
A_2	Les antécédents du répondant peuvent influencer l'apparition d'un catalyseur. (exemple du tabagisme comme facteur de risque du cancer du poumon)
A_3	Les antécédents peuvent influencer le niveau du construit que l'on cherche à mesurer à chaque temps de mesure. (Seul le chemin partant des antécédents vers la mesure au temps 1 est représenté, car l'effet au temps 2 est inclus dans le chemin TC1 4)
A_4	Les antécédents peuvent influencer les mécanismes déclenchés par le catalyseur. (exemple : face à un même catalyseur, deux individus avec des traits de personnalité différents ne mobiliseront pas les mêmes stratégies d'ajustement)
Chemins venant du construit que l'on cherche à mesurer aux deux temps de mesure	
$TC1_1$ $TC2_1$	Le construit que l'on cherche à mesurer explique (en partie) la mesure observée.
$TC1_2$	Le construit que l'on cherche à mesurer au temps 1 peut influencer l'apparition du catalyseur. (exemple d'un niveau élevé de fatigue [construit ciblé] pouvant provoquer un accident de voiture [catalyseur])
$TC1_3$	Le construit que l'on cherche à mesurer au temps 1 peut induire des mécanismes. (exemple d'un niveau élevé de fatigue [construit ciblé] pouvant induire une recherche de soutien [mécanismes])
$TC1_4$	Le construit que l'on cherche à mesurer au temps 1 influence, en partie, le construit que l'on cherche à mesurer au temps 2.
Chemins venant de la mesure obtenue au temps 1	
$Me1_1$	La mesure au temps 1 peut influencer la mesure au second temps. Ce chemin correspondrait à la corrélation entre les facteurs résiduels (c'est-à-dire tout ce qui est propre aux mesures et à la variation aléatoire de l'erreur).
Chemins venant du catalyseur	
C_1	Le catalyseur déclenche des mécanismes d'ajustement au changement d'état de santé.
C_2	Le catalyseur peut influencer le construit que l'on cherche à mesurer au temps 2. Il s'agit généralement de la question de recherche de nombreuses études (exemple du diagnostic d'une maladie affectant la qualité de vie).
C_3	Le catalyseur peut directement influencer les résultats de la mesure au temps 2. Si ce chemin survient, alors le changement observé ne peut pas être entièrement expliqué par le changement que l'on cherchait à mesurer : il y aura du *response shift*.
Chemins venant des mécanismes	
M_1	Les mécanismes peuvent influencer le niveau du construit que l'on cherche à mesurer au temps 2. (exemple, d'un individu qui ressent moins de fatigue au deuxième temps après avoir recherché du soutien)
M_2	Les mécanismes peuvent influencer les résultats de la mesure au temps 2. Si les chemins C_1 (effet du catalyseur sur les mécanismes) et M_2 (effet des mécanismes sur la mesure au temps 2) surviennent, alors le catalyseur a un impact sur la mesure au temps 2 par l'intermédiaire des mécanismes (chemins C_1 puis M_2). Dans ce cas, le changement observé ne sera pas entièrement expliqué par le changement que l'on cherchait à mesurer : il y aura du *response shift*.

L'une des implications majeures de cette nouvelle définition est que les 3R (Recalibration, Repriorisation et Reconceptualisation) n'apparaissent plus dans la définition du *response shift*. En effet, d'après Vanier *et al.* [18] ces concepts ne sont pas nécessairement du *response shift* en soi, ils expliquent plutôt comment le *response shift* peut survenir. Par exemple, l'interaction entre le catalyseur, les antécédents et les mécanismes peut pousser les individus à recalibrer leurs normes de mesure internes, revoir l'importance qu'ils accordent à certains domaines, ou encore reconceptualiser le construit ciblé par le questionnaire étudié, menant ainsi à une dissonance entre le changement observé et l'évolution du construit ciblé (c.-à-d., du *response shift*). De même, la notion d'*appraisal* a, elle aussi, été placée par Vanier *et al.* dans les théories expliquant comment le *response shift* survient. Une autre implication est la disparition du terme de "qualité de vie", le modèle du *response shift* pouvant s'appliquer à d'autres *Patient-Reported Outcomes*.

Ces auteurs rappellent néanmoins que des études futures seront nécessaires pour tester empiriquement les hypothèses faites et les relations qui ont été supposées. Aussi, ce modèle est potentiellement amené à évoluer [18].

3.3.3 Approches méthodologiques pour la détection du *response shift*

De nombreuses approches ont été développées pour détecter et prendre en compte le *response shift* lors de l'analyse de données longitudinales [23, 112–115]. Ces méthodes peuvent être scindées en deux grandes familles [113] :

- Les méthodes qui nécessitent un design d'étude ou un outil de mesure spécifique. Ce type de méthodes est utilisé lorsque la survenue de *response shift* est anticipée et qu'elle constitue un critère d'intérêt.
- Les méthodes statistiques qui cherchent à inférer du *response shift* à partir de données déjà recueillies.

Les méthodes présentées opérationnalisent toutes la définition du *response shift* proposée par Sprangers et Schwartz [17], c'est-à-dire qu'elles recherchent la présence d'une ou plusieurs "forme(s)" de *response shift* (recalibration, repriorisation ou reconceptualisation). De plus, certaines méthodes permettent également d'étudier l'écart entre le changement de score observé et l'évolution du construit qui était ciblé, opérationnalisant ainsi la définition de Oort [110] (qui a depuis été clarifiée et précisée par Vanier *et al.* [18])

Les méthodes basées sur un design d'étude ou un outil de mesure spécifique

L'approche then-test

Cette approche est basée sur un design d'étude particulier puisqu'elle nécessite d'obtenir trois évaluations du construit étudié [102, 116] :

- Une évaluation obtenue avant l'événement supposé induire des changements dans la signification de l'autoévaluation du construit mesuré (mesure *pré-test*) ;
- Une évaluation obtenue après l'événement (mesure *post-test*) ;
- Une évaluation rétrospective où l'on demande à l'individu de réévaluer son niveau initial avant l'événement (mesure *then-test*).

Les évaluations *post-test* et *then-test* doivent avoir lieu au même moment pour s'assurer que les normes internes mobilisées par l'individu lors de l'évaluation soient les mêmes. L'écart entre les évaluations *pré-test* et *then-test* est supposé être une estimation de l'ampleur de l'effet du *response shift* (recalibration). La différence entre les évaluations *post-test* et *then-test* est utilisé comme mesure du changement ajusté sur le *response shift*. Avec cette méthode, il est possible de discerner dans quelle mesure le changement de score observé représente l'évolution du construit étudié et dans quelle mesure il reflète du *response shift*. Cette approche est représentée graphiquement par la figure 3.11. Il s'agit d'une approche conçue pour rechercher de la recalibration au niveau de l'échantillon (comparaison des moyennes *pré-test* et *then-test*), mais elle peut être adaptée pour réaliser des analyses en sous-groupes [23]. Bien que très répandue pour évaluer l'effet de la recalibration, son utilisation présente plusieurs inconvénients. Des chercheurs ont notamment remis en question le principe de base de l'approche, selon lequel les normes internes des individus sont les mêmes lors des évaluations *post-test* et *then-test* [117]. De plus, des phénomènes autres que la recalibration peuvent également entraîner des différences entre les moyennes *pré-test* et *then-test*. Parmi les explications alternatives possibles, on peut par exemple nommer le biais de désirabilité sociale ou le biais de mémoire (du fait de la nature rétrospective de l'évaluation *then-test*) [23, 102, 118].

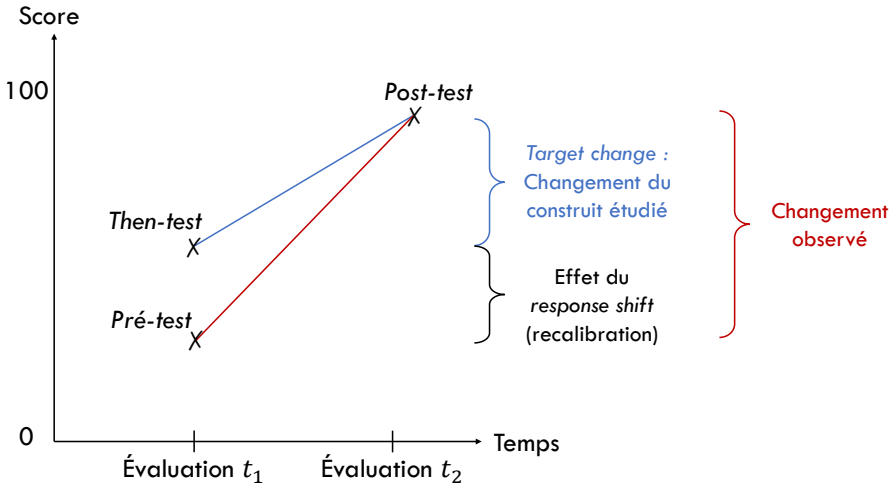


FIGURE 3.11 – Représentation graphique de l'approche *then-test*

PGI et SEIQoL : des questionnaires de qualité de vie individualisés

Les questionnaires PGI (*Patient Generated Index*) [119] et SEIQoL (*Schedule for the evaluation of Individual Quality of Life*) [120] sont deux questionnaires de qualité de vie individualisés où il est demandé au répondant d'identifier cinq domaines importants pour sa qualité de vie, d'évaluer son niveau de fonctionnement pour chacun de ces domaines et d'indiquer le niveau d'importance relative qu'il accorde à chaque domaine cité en tant que déterminant de sa qualité de vie[3]. Ces mesures de qualité de vie individualisées permettent d'investiguer le *response shift* au niveau de l'échantillon et au niveau individuel en examinant les changements au cours du temps dans les domaines sélectionnés (reconceptualisation) et dans l'importance accordée aux domaines mentionnés (repriorisation) [23, 122, 126, 127].

L'approche par vignettes

Afin de détecter du *response shift*, Korage *et al.* [128] ont proposé en 2007 d'utiliser un nouvel outil : les vignettes. Les vignettes sont de petits textes décrivant en quelques lignes différents états de santé hypothétiques. Par exemple : "Monsieur A n'a aucun problème pour se déplacer, se laver ou s'habiller, a des fuites urinaires quotidiennes, n'a aucune douleur ou gêne, n'est pas anxieux ou déprimé" (vignette issue de Korage *et al.* [128], librement traduite). Korfage *et al.* ont proposé cette approche pour rechercher du *response shift* (repriorisation) à la suite d'un diagnostic de cancer de la prostate. Leur étude consistait à faire évaluer trois vignettes à l'aide d'une échelle visuelle analogique (0 = état de santé "très mauvais" et 10 = état de santé "très bon") à des hommes à trois temps de mesure différents : t_1 = avant le diagnostic du cancer, t_2 = un mois après le diagnostic (avant l'initiation des traitements) et t_3 = sept mois après le diagnostic (après l'initiation des traitements). Leur hypothèse était que les hommes atteints d'un

3. Pour le PGI, deux versions de cette étape existent. Dans la première version, il est demandé aux patients d'indiquer les domaines qu'ils voudraient le plus améliorer et ceux qui ne sont pas si importants que cela [119] (on parle alors d'importance relative d'une amélioration potentielle). Dans la seconde version, il est cette fois-ci demandé aux répondants d'indiquer les domaines importants pour leur qualité de vie [121]. Les travaux pour investiguer le *response shift* à l'aide du PGI se sont appuyés sur la première version (on peut par exemple citer les travaux d'Ahmed *et al.* [122, 123] et d'Aburub *et al.* [124]). L'adaptation française du questionnaire proposée par Botella *et al.* s'appuie en revanche sur la seconde version [125].

cancer de la prostate pourraient considérer les états de santé décrits par les vignettes comme moins préjudiciables suite au diagnostic de leur cancer (en effet, après le diagnostic, ces patients savent qu'ils risquent de connaître eux-mêmes les dysfonctionnements décrits). Par la suite, les vignettes ont été utilisées par Hinz *et al.* [129–131] et Preiss *et al.* [132] afin de rechercher du *response shift* dans un cadre transversal. Dans leurs articles, ces auteurs définissent le *response shift* comme des changements dans le système de référence ("*frame of reference*") des patients, sans qu'un type de changement précis ne soit mentionné (reconceptualisation ? repriorisation ? recalibration ? appraisal ?). Ces auteurs s'attendaient à ce que les patients malades évaluent un même état de santé plus favorablement que les individus contrôles non malades. Enfin, dans un contexte longitudinal, des vignettes construites sur la base du questionnaire SF-12 [133] ont également été utilisées par Topp *et al.* [134]. Ces auteurs ont néanmoins fait face à de nombreux défis au cours de leur étude, ce qui les a amenés à s'interroger sur la pertinence de l'approche par vignettes pour la recherche du *response shift*[4]. Il est à noter que l'approche par vignettes est conçue pour détecter du *response shift* au niveau de l'échantillon (comparaison des moyennes des évaluations vignettes), mais elle peut être adaptée pour réaliser des analyses en sous-groupes [23].

Les entretiens semi-dirigés

Il s'agit d'une approche individuelle qui consiste à interroger directement les personnes au cours d'entretiens qualitatifs. Les questions de ces entretiens visent à amener les personnes interrogées à verbaliser un potentiel *response shift*. Ces questions permettent de détecter :

- De la recalibration

  Exemple : Auriez-vous évalué le niveau de votre qualité de vie de la même manière avant votre accident si on vous l'avait demandé à ce moment-là plutôt que maintenant (rétrospectivement) ? Est-ce que la modalité de réponse [citer la modalité de réponse] a toujours eu la même signification pour vous ?

- De la repriorisation

4. Ces auteurs ont remis en question la fiabilité *test-restest* de leurs vignettes. En effet, les réponses des individus aux vignettes fluctuaient de façon non directionnelle avec une faible concordance intra-individuelle. De plus, l'interprétation des données qualitatives obtenues grâce à des entretiens de type *think aloud* s'est révélée difficile.

Exemple : Y a-t-il des choses qui sont devenues plus/moins importantes pour vous maintenant ?

– De la reconceptualisation

Exemple : est-ce que la signification de la qualité de vie a changé pour vous ? Y a-t-il des choses qui sont importantes aujourd'hui pour votre qualité de vie et qui ne l'étaient pas du tout avant (ou au contraire des choses qui n'ont plus du tout d'importance) ?

Les questions prises en exemples proviennent des travaux de Beeken *et al.* [135], elles ont été librement traduites et adaptées.

Une méthode assez proche des entretiens semi-dirigés est l'approche *think aloud* (méthode de la pensée à voix haute). Il s'agit d'entretiens visant à expliciter les processus cognitifs à l'œuvre lors de la passation d'un questionnaire. Au cours de ces entretiens, il est demandé aux participants de répondre à un questionnaire tout en explicitant à l'oral la façon dont ils interprètent les items qui le composent. Cette approche peut permettre de mettre en évidence des changements dans l'interprétation des items ou du construit au cours du temps.

QoLAP : une mesure de l'appraisal

Le QoLAP (*Quality of Life Appraisal Profile*) a été développé par Rapkin et Schwartz dans l'objectif d'évaluer l'*appraisal* (les processus cognitifs impliqués lors de la passation d'un questionnaire) [108]. Ce questionnaire est conçu pour être utilisé en parallèle d'un outil de mesure de la qualité de vie (les instructions supposent que le patient vient juste de compléter de tels questionnaires). Il contient des items qui permettent d'interroger le répondant sur son cadre de référence, cerner les stratégies implicites qu'il utilise pour échantillonner ses expériences, s'intéresser aux normes de comparaison considérées lors de l'évaluation de la qualité de vie, et enfin, chercher à comprendre la pondération utilisée par le répondant. Bien que le questionnaire soit individuel, les analyses proposées par Rapkin et Schwartz sont au niveau de l'échantillon (modèle de régression multivarié où la variable à expliquer est le changement de score de qualité de vie et les variables explicatives sont les influences "standard" et les changements dans les scores d'*appraisal*). Du *response shift* est détecté quand les changements dans les scores d'*apparaisal* permettent d'expliquer une part de variance de l'évolution observée de la qualité de vie, qui n'est

pas déjà expliquée par les influences standard. La traduction de ces effets en termes de recalibration, repriorisation et reconceptualisation n'a pas encore été proposée par ces auteurs. D'autres mesures de l'*appraisal* ont été proposées par Rapkin, Schwartz et leurs co-auteurs : le QoLAP version 2 [136] et le *Brief Appraisal Inventory* [137]. Ces questionnaires ont été récemment critiqués par Verdam et Oort [138]. En effet, la représentation de l'*appraisal* en tant que variable leur semble difficile à envisager : que signifierait le fait d'avoir "plus" ou "moins" d'*appraisal* ? Avoir des scores élevés ou faibles pour "le cadre de référence", " les normes de comparaison", "les stratégies de sélection d'expériences" ou "la pondération" ? Ainsi, d'après ces auteurs, la représentation quantitative des processus cognitifs à l'œuvre lors de la passation des questionnaires semble poser des problèmes d'interprétation et ne permet donc pas d'étudier le rôle de l'*appraisal* dans la survenue du *response shift*.

Les méthodes statistiques

La procédure de Oort

En 2005, Oort [139] a proposé une méthode de détection statistique du *response shift* entre deux temps de mesure, basée sur les modèles à équations structurelles (SEM), où les variables manifestes sont les scores aux dimensions du questionnaire. En utilisant les SEM, Oort a opérationnalisé les différentes formes de *response shift* (recalibration, repriorisation et reconceptualisation) par des changements au cours du temps dans les valeurs des paramètres du modèle de mesure. Ainsi, la détection du *response shift* est ramenée à la recherche de paramètres non-invariants :

- Un changement dans le *pattern* des *factor loadings* reliant les variables latentes ciblées par le questionnaire aux scores des dimensions qu'elles sont censées influer suggère de la **reconceptualisation**. Ce type de changement survient par exemple lorsque le *factor loading* reliant une variable latente au score d'une dimension est nul au premier temps de mesure, mais non nul au second temps de mesure. Cela peut alors signifier que la dimension ne reflétait initialement pas la variable latente, mais que cela a changé par la suite. De façon similaire, un *factor loading* peut être non nul au départ et devenir nul au second temps. Cela peut signifier que la dimension reflétait initialement la variable latente, mais qu'elle ne la reflète plus par la suite.
- Les autres types de changement dans les *factor loadings* suggèrent de la **repriorisation**. En effet, si la valeur d'un *factor loading* reliant une variable latente au score d'une dimension change au cours du temps, cela peut signifier que la dimension change d'importance pour la mesure du construit étudié.
- Des changements dans les *intercepts* suggèrent de la **recalibration uniforme**. En effet, d'après Oort, si les répondants changent leur interprétation des modalités de réponse, et si ce changement affecte toutes les modalités de réponse de la même façon (direction et magnitude) cela devrait se manifester par un changement dans les moyennes des scores, et donc un changement au niveau des *intercepts*. Si les changements ne vont pas dans la même

direction ou varie en magnitude, cela se pourrait se manifester sur les variances résiduelles (mais aussi éventuellement sur les *intercepts*). Ainsi, des changements dans les variances résiduelles suggèrent de la **recalibration non uniforme**.

En pratique, la procédure de Oort se déroule en 4 étapes représentées dans la figure 3.12. Chacune de ces étapes est associée à un modèle spécifique. Ces étapes sont décrites ci-dessous :

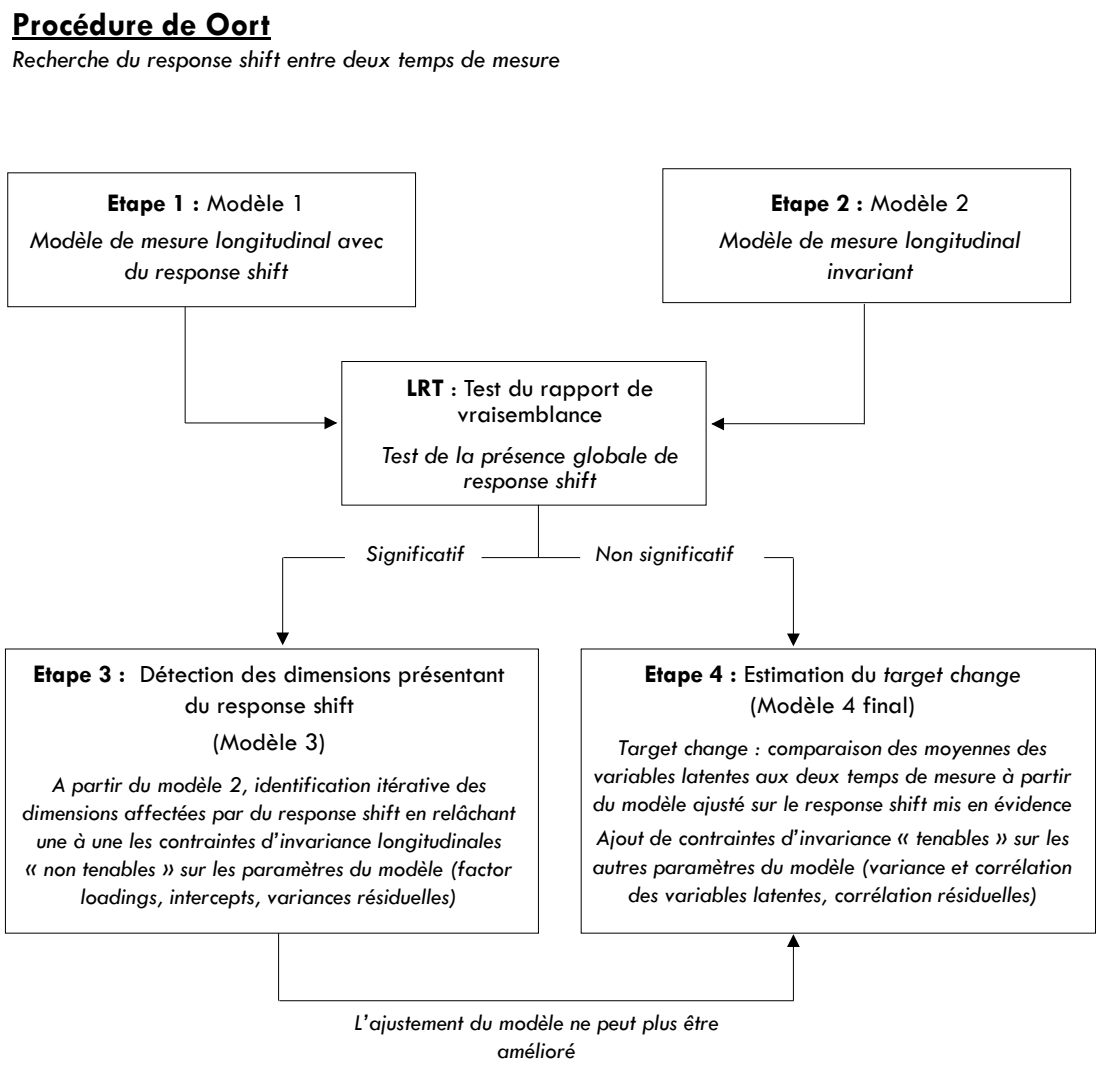


FIGURE 3.12 – Représentation schématique de la procédure de Oort

Étape 1 : Établir un modèle de mesure (modèle 1)

Il s'agit de spécifier un modèle de mesure longitudinal pour le SEM, défini en fonction de la structure du questionnaire étudié. Dans ce modèle, à chaque temps de mesure, les variables latentes (représentant les construits multidimensionnels ciblés par le questionnaire) sont reliées par un lien linéaire aux scores des dimensions qu'elles sont censées influer. Aucune contrainte d'invariance longitudinale n'est imposée sur les *factor loadings*, les *intercepts* et les variances résiduelles : ces paramètres sont donc librement estimés à chaque temps. On parle alors de "modèle non-invariant" (ou "modèle avec du *response shift*"). Pour que le modèle soit identifiable, les moyennes et les variances des variables latentes sont respectivement fixées à 0 et 1 aux deux temps de mesure. En cas d'ajustement insatisfaisant du modèle, il est possible de le re-spécifier, en ajoutant par exemple des covariances entre les termes d'erreur. La re-spécification du modèle doit avoir du sens : elle doit pouvoir se justifier. S'il n'est pas possible d'identifier un modèle présentant un bon ajustement, cela peut signifier qu'il y a eu de la reconceptualisation, et que cette reconceptualisation est si importante qu'il n'est pas possible d'identifier un modèle adéquat. Dans ce cas, la procédure s'arrête ici. Afin d'illustrer cette étape, Oort [139] a proposé un modèle de mesure longitudinal possible pour les dimensions du questionnaire SF-36 représenté en figure 3.13.

Étape 2 : Tester la présence globale de *response shift*

Afin de tester la présence globale de *response shift*, un second modèle (modèle 2) est estimé. Dans ce modèle, les *factor loadings*, les *intercepts* et les variances résiduelles sont contraints à être constants aux deux temps de mesure (invariance longitudinale). On parle de "modèle invariant" (ou "modèle sans *response shift*"). L'ajustement des deux modèles est comparé à l'aide d'un test du rapport de vraisemblance (*Likelihood-Ratio Test*, LRT). Si le test est significatif, l'ajustement du modèle 1 est meilleur, suggérant la présence de *response shift*, la procédure se poursuit avec l'étape 3 (identification des dimensions présentant du *response shift*). Au contraire, un test non significatif suggère l'absence de *response shift*. La procédure continue dans ce cas avec l'étape 4.

Étape 3 : Détection des dimensions présentant du *response shift*

Lors de l'étape 3, le *response shift* est détecté en déterminant si l'ajustement du modèle 2 (modèle sans *response shift*) peut être significativement amélioré en libérant certaines contraintes d'invariance longitudinales portant sur les *factor loadings*, les *intercepts* et les variances résiduelles. Il s'agit en fait d'une étape itérative, dont le point de départ est le modèle 2, et où les contraintes d'invariance "non tenables" sont libérées une par une. À chaque fois qu'une contrainte est libérée et que le modèle est ré-estimé, l'amélioration de l'ajustement du modèle est testée. Si l'ajustement du modèle est meilleur une fois la contrainte libérée, le modèle est alors mis à jour. L'étape se poursuit jusqu'à ce que l'ajustement du modèle ne puisse plus être amélioré, l'étape arrive alors à sa fin, le modèle final obtenu (modèle 3) prend en compte toutes les formes de *reponse shift* détectées (associées aux contraintes d'invariance libérées).

On peut remarquer qu'à chaque itération, la contrainte libérée peut porter sur n'importe quelle dimension et être associée à n'importe quelle forme de *response shift* : il n'y a pas de hiérarchie dans la recherche [5].

Au cours de cette étape, la recherche des contraintes non tenables peut par exemple être guidée par les indices de modification. Néanmoins, les libérations des contraintes effectuées doivent faire sens : elles doivent pouvoir être justifiées théoriquement.

Étape 4 : Modèle final - Estimation du *target change*

À partir du modèle 3 (si le LRT était significatif) ou à partir du modèle 2 (si le LRT était non significatif), les moyennes des variables latentes peuvent être comparées entre les deux temps de mesure. Si l'hypothèse nulle d'égalité des moyennes entre les deux temps de mesure est rejetée, alors l'évolution des moyennes des variables latentes peut être considérée comme une mesure du *target change* (évolution du construit ciblé par le questionnaire, ajusté sur le *response shift* mis en évidence). D'autres changements peuvent être étudiés au cours de cette étape comme l'évolution : des variances des variables latentes, des corrélations entre variables latentes et des

5. Pour pouvoir détecter de la reconceptualisation, il peut être nécessaire d'introduire des *cross-loadings*

corrélations entre les résidus. Suite à la réalisation de tests statistiques, des contraintes d'invariance longitudinales tenables sur ces paramètres peuvent être ajoutées une à une au modèle 3, menant au modèle 4 (modèle final).

Bilan et extensions : Pour résumer, cette méthode permet de déterminer si une ou plusieurs dimensions d'un questionnaire sont affectées par du *response shift*, préciser la ou les formes de *response shift* impliqué(e)s et quantifier leur ampleur. La détection du *response shift* est basée sur l'estimation en cascade de modèles à équations structurelles. On parle de détection du *response shift* au niveau des dimensions (*domain-level*) car l'analyse a lieu au niveau des scores des dimensions. La procédure de Oort permet également d'estimer le changement au niveau des variables latentes ciblées par le questionnaire en prenant en compte le *response shift* détecté (ce changement est désigné par le terme *target change*). Les contributions respectives du *target change* et du *response shift* dans le changement de score observé pour chacune des dimensions peuvent également être décrites. La détection du *response shift* a lieu au niveau de l'échantillon : elle suppose que le *response shift* est un phénomène homogène (c'est-à-dire qu'une majorité d'individus de l'échantillon l'expérimentent de la même façon). Elle suppose également que le *response shift* touche une minorité de variables manifestes (ici les scores des dimensions). Depuis, des extensions de la procédure de Oort ont été réalisées, notamment pour :

– Considérer plus de deux temps de mesure pour mieux appréhender la temporalité du *response shift* [113, 140, 141].

– Étudier les variations interindividuelles du *response shift* en : (i) incluant des covariables représentant des caractéristiques de patients dans le SEM [142] ou (ii) en stratifiant l'analyse [143].

– Étudier le *reponse shift* au niveau des items [144, 145].

Une adaptation de la procédure a également été proposée par Nolte *et al.* [117] afin d'introduire une hiérarchie dans la recherche du *response shift*.

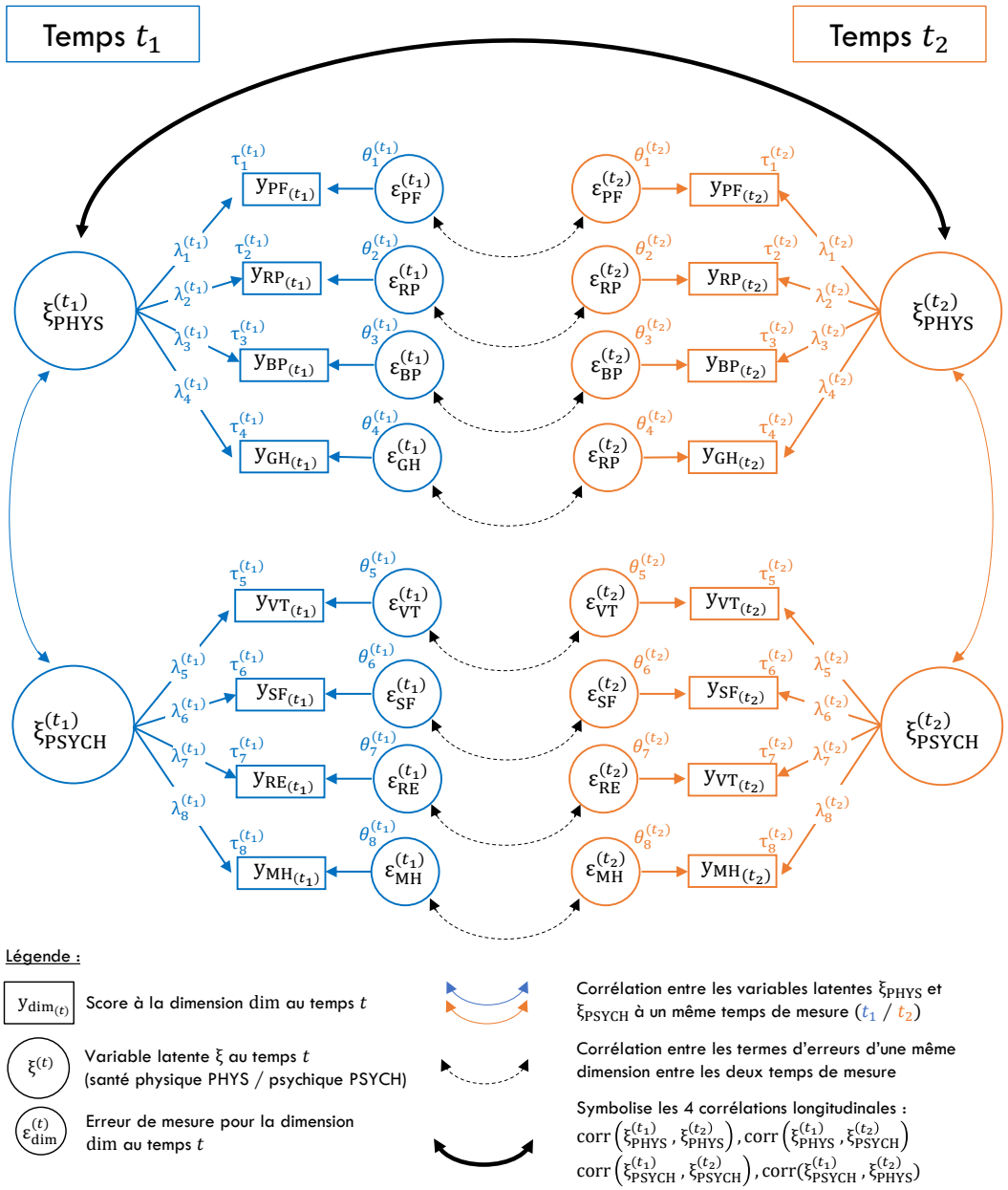


FIGURE 3.13 – Représentation graphique du modèle de mesure longitudinal pour le questionnaire de qualité de vie liée à la santé SF-36

Notes : Les carrés représentent les variables observées (les scores aux dimensions du SF-36, notés y), les cercles ξ_{PHYS} et ξ_{PSYCH} représentent les variables latentes sous-jacentes (santé physique et psychique) et les cercles ε_{PF}, ..., ε_{MH} représentent les résidus.

L'algorithme ROSALI

Depuis quelques années, l'intérêt de la détection du *response shift* au niveau des items s'est accru [144–149]. En effet, les méthodes de détection au niveau des dimensions ne permettent pas de déterminer les items à l'origine de l'effet observé. De plus, ces analyses peuvent ne pas refléter correctement ce qu'il se passe au niveau des items, en particulier si le *response shift* a des effets opposés sur différents items. Ainsi, une détection du *response shift* au niveau des items pourrait apporter des informations complémentaires.

Pour répondre à ce besoin, l'algorithme ROSALI-IRT (RespOnse Shift ALgorithm at the Item level based on IRT) a été proposé par Guilleux *et al.* [147] afin de mener la détection du *response shift* au niveau des items entre deux temps de mesure en se basant sur la théorie de réponse à l'item (théories de la mesure de Rasch et de Lord). Cet algorithme reprend la structure de la procédure de Oort [139] et permet de rechercher, au sein des items d'une dimension, de la recalibration puis de la repriorisation à l'aide d'un GPCM longitudinal estimé sous contraintes :
– Un changement dans les paramètres de seuil des items indique de la recalibration (uniforme ou non) ;
– Un changement dans le paramètre de discrimination des items indique de la repriorisation.

La version proposée dans l'article de Guilleux *et al.* [147] a récemment fait l'objet de critiques[6], elle ne sera donc pas présentée dans ce manuscrit. À la place, une autre version de l'algorithme ROSALI, basée cette fois-ci sur la théorie de la mesure de Rasch (ROSALI-RMT), est présentée. Cette version a été proposée par Blanchin *et al.* [149] et est toujours inspirée de la procédure de Oort [139]. Elle permet de rechercher de la recalibration au sein d'une dimension à l'aide d'un PCM longitudinal estimé sous contraintes. Au cours de cet algorithme, l'identification d'un changement dans les paramètres de seuil de certains items indique de la recalibration (uniforme ou non).

6. Ces critiques portaient notamment sur : (i) l'étape préliminaire d'estimation des paramètres des items (estimations auxquelles l'algorithme se rapportait par la suite en négligeant l'incertitude qui leur était associée) (ii) l'absence de critère d'arrêt pour l'étape itérative.

En pratique, la procédure se déroule en 4 étapes représentées dans la figure 3.14. Chacune de ces étapes est associée à un PCM longitudinal spécifique. Ces étapes sont décrites ci-dessous :

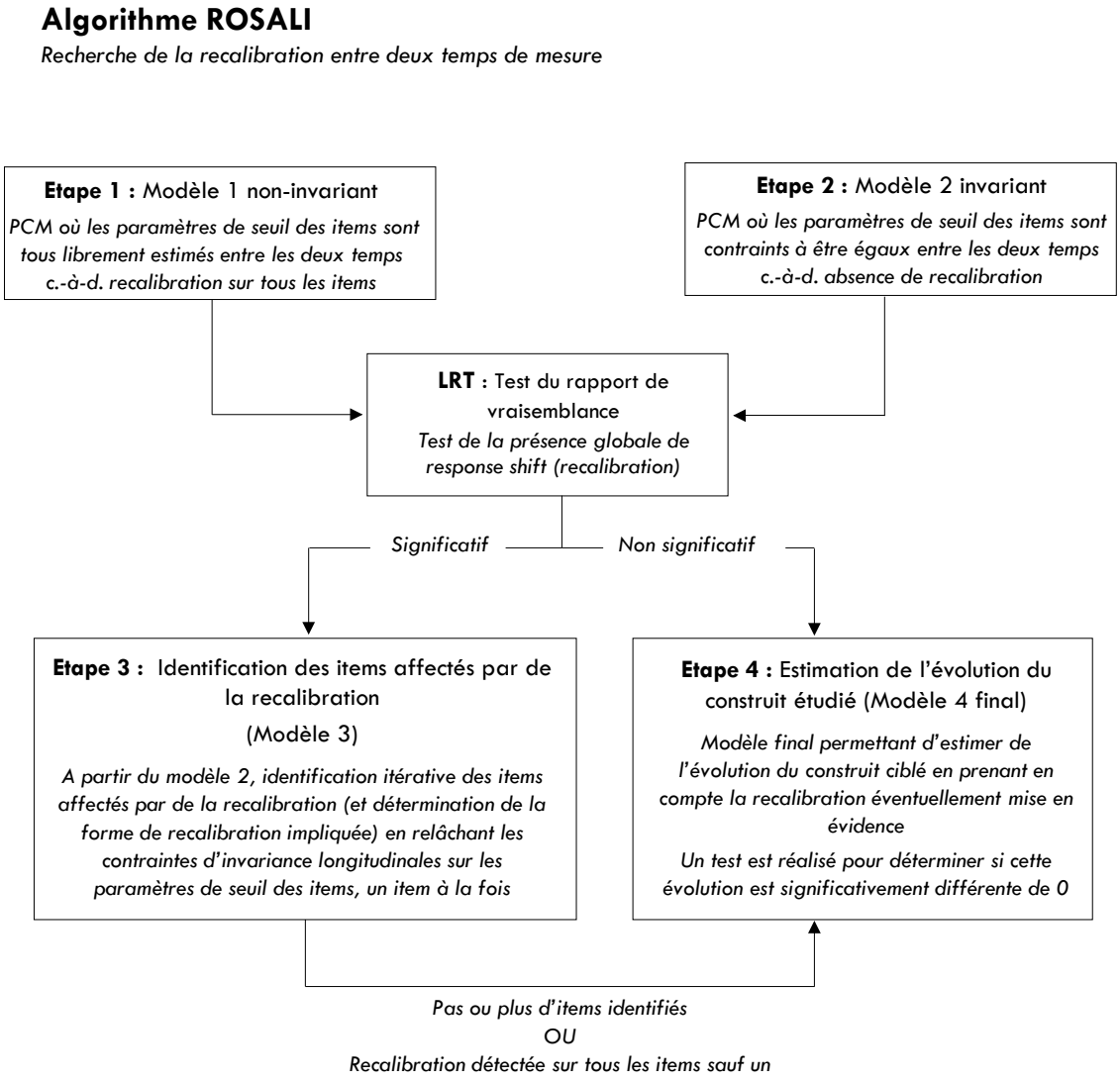


FIGURE 3.14 – Représentation schématique de l'algorithme ROSALI-RMT

Étape 1 : PCM longitudinal avec de la recalibration (modèle 1)

Dans cette première étape, un premier PCM longitudinal (modèle 1) supposant de la recalibration sur tous les items est estimé. Dans ce modèle, aucune contrainte d'invariance n'est imposée sur les paramètres de seuil des items : ils sont donc librement estimés entre les deux temps de mesure. Pour que le modèle soit identifiable, la moyenne de la variable latente est fixée à zéro aux deux temps.

Étape 2 : PCM longitudinal sans recalibration (modèle 2) et test de la présence globale de la recalibration

Afin de tester la présence globale de recalibration, un second PCM longitudinal (modèle 2) est estimé. Dans ce modèle, les paramètres de seuil des items sont contraints à être constants aux deux temps de mesure (invariance longitudinale). Ce modèle est comparé au modèle 1 en utilisant un test du rapport de vraisemblance (LRT). En cas de test significatif, l'algorithme passe à l'étape suivante (étape 3) pour identifier les items sur lesquels la recalibration se manifeste. Dans le cas contraire (test non significatif), on suppose qu'il n'y a pas de recalibration, le modèle 2 est retenu et tous les paramètres d'items sont supposés être invariants (égaux entre les deux temps de mesure) : l'algorithme passe alors à l'étape 4 où le modèle final correspond au modèle 2.

Étape 3 : Détection des items avec de la recalibration

L'étape 3 est une étape itérative qui a pour objectif d'identifier les items affectés par de la recalibration, ainsi que la forme de recalibration sous-jacente. Un nouveau modèle, appelé modèle 3, est introduit de sorte que le modèle 3 corresponde au modèle 2 (modèle sans recalibration) au début de l'étape.

– À partir du modèle 3, l'algorithme estime pour chaque item un nouveau modèle où la contrainte d'invariance longitudinale associée à l'item étudié est relâchée (estimation libre des paramètres de seuil associés à cet item entre les deux temps de mesure) et où les contraintes d'invariance pour les autres items restent inchangées. À partir de ces nouveaux modèles, un test statistique

est réalisé pour déterminer si les paramètres de seuil de l'item étudié diffèrent significativement entre les deux temps de mesure.

– L'algorithme sélectionne ensuite le modèle associé au test le plus significatif après une correction pour la multiplicité des tests (correction de Bonferroni). Si aucun test n'est significatif, alors l'algorithme passe directement à l'étape 4. Sinon, on suppose alors que l'item associé au modèle sélectionné (l'item pour lequel la contrainte d'invariance longitudinale a été relâchée) est affecté par de la recalibration. La forme de la recalibration est déterminée à l'aide d'un nouveau test statistique : la recalibration est dite "uniforme" si les paramètres de seuil évoluent tous de la même façon (même direction et magnitude). Elle est dite "non uniforme" dans le cas contraire.

– Le modèle 3 est ensuite mis à jour pour prendre en compte la forme de recalibration mise en évidence.
L'item sélectionné ne sera plus testé. L'étape 3 est répétée sur tous les items restants. Elle s'arrête lorsqu'il n'y a plus d'item associé à un test significatif ou juste avant de libérer le dernier item invariant.

Étape 4 : Estimation de l'évolution du construit étudié

Cette dernière étape permet d'estimer l'effet temps sur le niveau de la variable latente, ajusté sur la recalibration précédemment mise en évidence grâce à un modèle final (modèle 4). Le modèle 4 correspond à la dernière version du modèle 3 si le test du rapport de vraisemblance était significatif. Sinon, il correspond au modèle 2.

Bilan et extensions : Pour résumer, cet algorithme permet de détecter de la recalibration au sein des items d'une dimension issue d'un questionnaire recueilli à deux temps de mesure. La recalibration détectée peut être uniforme ou non uniforme. L'algorithme ROSALI est basé sur l'estimation en cascade de PCM longitudinaux. On parle de détection de la recalibration au niveau des items (*item-level*) car l'analyse permet d'identifier les items sur lesquels la recalibration se manifeste. Cet algorithme permet également d'estimer l'évolution du construit étudié en

prenant en compte la recalibration détectée. Tout comme la procédure de Oort, cette méthode suppose que la recalibration est un phénomène homogène qui ne touche qu'une minorité d'items.

Une extension de l'algorithme ROSALI a été proposé par Hammas *et al.* [150] pour pouvoir étudier les variations interindividuelles de la recalibration en incluant une covariable binaire dans l'analyse (représentant une caractéristique des patients). Cette extension comporte deux parties :

- Partie 1 : Cette première partie vise à étudier l'invariance de la mesure entre groupes au premier temps de mesure. L'objectif est de déterminer s'il existe initialement des différences dans les paramètres d'items entre les deux groupes. Cette détection est basée sur l'estimation en cascade de PCM transversaux. Elle se déroule en 4 étapes décrites en figure 3.15.
- Partie 2 : La seconde partie permet de détecter de la recalibration entre les deux temps de mesure (en prenant en compte les différences initiales dans les paramètres d'items qui ont été mises en évidence à l'étape 1). La détection de la recalibration s'effectue ensuite en 4 étapes à l'aide de l'estimation de PCM longitudinaux en cascades (de façon similaire à l'algorithme ROSALI sans covariable). Une différence peut néanmoins être soulignée à l'étape 3 : lorsqu'un item semble être affecté par de la recalibration, l'algorithme cherche d'abord à déterminer si la recalibration est la même dans les deux groupes (recalibration "commune") ou non (recalibration "différentielle"). Si de la recalibration commune est mise en évidence, l'algorithme cherche ensuite à déterminer sa forme. Si de la recalibration différentielle est mise en évidence, l'algorithme détermine séparément pour chacun des deux groupes si de la recalibration survient et, dans l'affirmative, quelle est sa forme. Cette seconde partie est représentée schématiquement dans la figure 3.16.

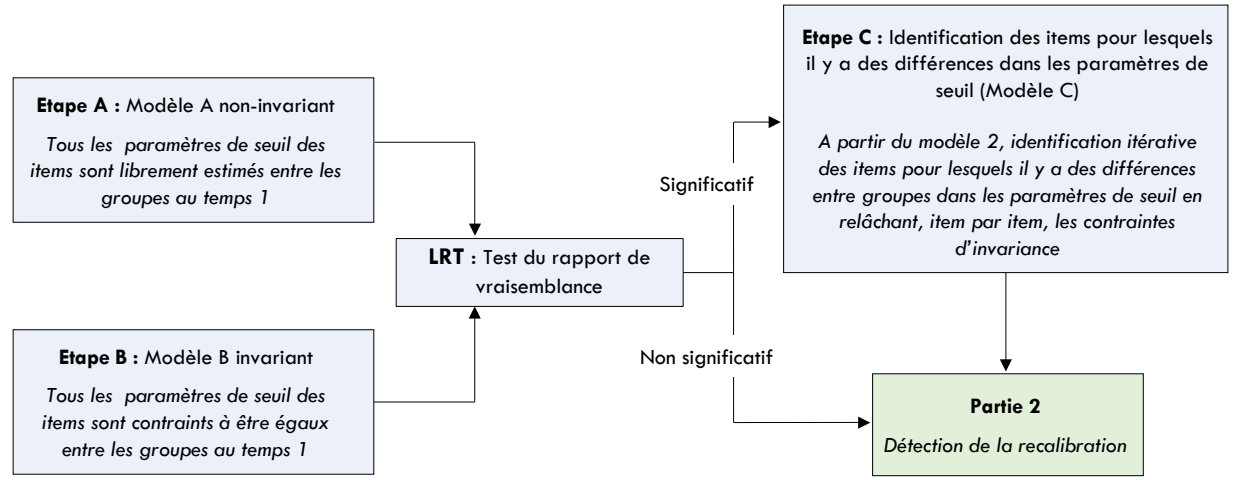


FIGURE 3.15 – Partie 1 de l'algorithme ROSALI avec une covariable
Source : Adapté et librement traduit de Blanchin et al., Methods, 2022

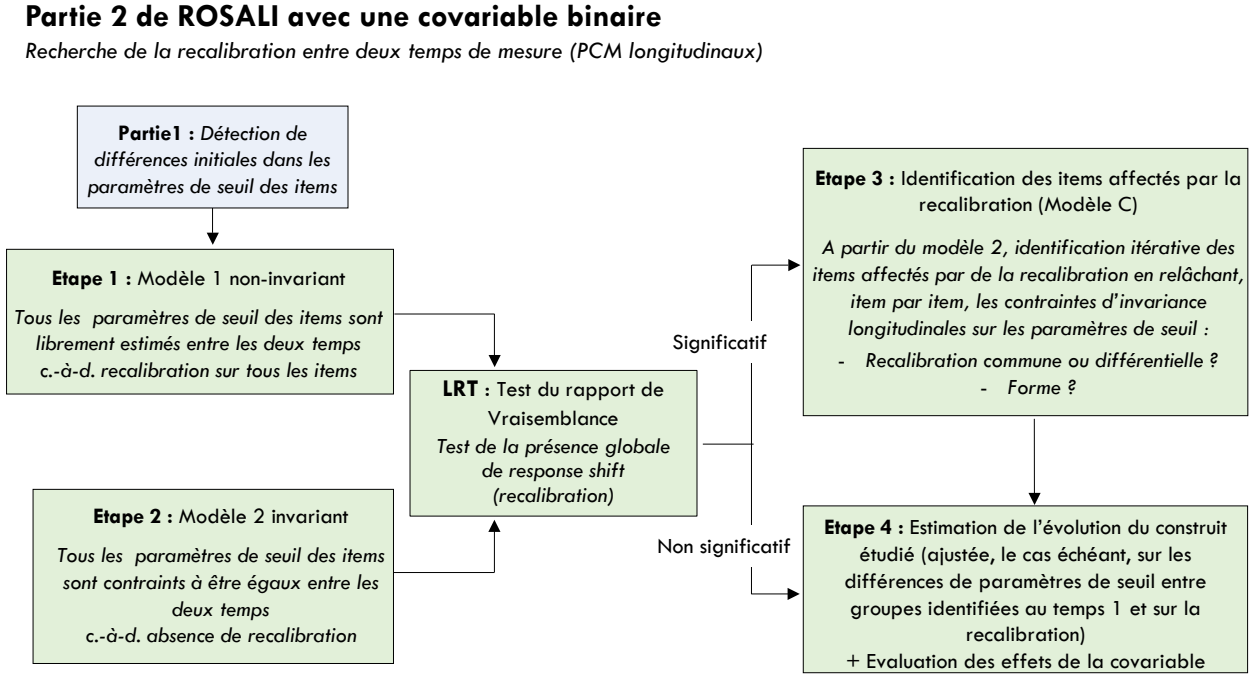


FIGURE 3.16 – Partie 2 de l'algorithme ROSALI avec une covariable binaire
Source : Adapté et librement traduit de Blanchin et al., Methods, 2022

Quelques remarques peuvent être faites à propos de cet algorithme. Premièrement, les modèles de la famille de Rasch ont été plébiscités par Blanchin *et al.* [149] parce qu'ils bénéficient de la propriété d'objectivité spécifique, ce qui signifie qu'une estimation consistante du niveau de la variable latente pourrait être obtenue même si certains items sont manquants (que le processus à l'origine des données manquantes soit ignorable ou non). De plus, dans un cadre transversal, la partie 1 de cette extension peut être vue comme une recherche de fonctionnement différentiel des items (DIF) entre les deux groupes d'individus définis par la covariable binaire introduite dans l'analyse. En effet, au cours de cette première partie, l'algorithme recherche des différences entre groupes dans les paramètres de seuil des items (différences pouvant être constantes ou non : DIF homogène ou non). Néanmoins, dans le cadre longitudinal de ROSALI, Hammas *et al.* ont décidé de dédier la terminologie de "DIF" à des différences entre groupes dans les paramètres de seuil des items qui se maintiendraient dans le temps [150].

L'étude de la trajectoire des résidus d'un modèle linéaire mixte

Cette méthode a été proposée par Mayo *et al.* [151] en 2008 afin de détecter le *response shift* (au sens large) à un niveau individuel. Cette méthode nécessite d'avoir au moins trois temps de mesure du construit étudié et peut se décomposer en trois étapes (décrites ci-dessous) :

Étape 1 : Construire un modèle pour prédire la mesure longitudinale du construit étudié

La première étape consiste à spécifier un modèle longitudinal à l'aide d'un modèle linéaire mixte à *intercept* aléatoire. Dans ce modèle, la variable à prédire Y est la mesure du construit étudié (par exemple un score de qualité de vie globale) et où les variables prédictives X_1, ..., X_p sont :

- Des variables recueillies à l'inclusion, comme le sexe, l'âge ou la sévérité de l'événement considéré comme catalyseur ;
- Des covariables dépendantes du temps (potentiellement issues de questionnaires autorappor-

tés).

Au départ, toutes ces variables sont introduites seules et en possible interaction avec le temps écoulé depuis l'événement catalyseur [7]. Seules les variables prédictives et les interactions associées à un effet significatif sont conservées dans le modèle. Mayo *et al.* indiquent par ailleurs qu'il ne faut pas inclure la variable "temps écoulé depuis le catalyseur" directement dans le modèle puisque cela pourrait compromettre la détection du *response shift*. En effet, les changements non expliqués du construit étudié qui sont dus au temps font partie de la définition du *response shift*. De plus, les variables prédictives doivent couvrir tous les domaines implicites du construit étudié et ne doivent pas être affectées par du *response shift*.

Étape 2 : Extraire du modèle prédictif les résidus centrés

Cette seconde étape vise à obtenir les résidus centrés pour chaque individu et chaque temps de mesure. Pour chaque individu i, et à chaque temps de mesure t, le résidu $r_i^{(t)}$ est obtenu en calculant la différence entre la mesure observée du construit étudié $y_i^{(t)}$ et la mesure prédite par le modèle $\hat{y}_i^{(t)}$:

$$r_i^{(t)} = y_i^{(t)} - \hat{y}_i^{(t)}$$

Les résidus obtenus à chaque temps de mesure pour un même individu i sont ensuite centrés en soustrayant la moyenne des résidus de l'individu étudié $\bar{r}_i$ (calculée sur l'ensemble des temps de mesure) :

$$\tilde{r}_i^{(t)} = r_i^{(t)} - \bar{r}_i$$

Ce centrage sur la moyenne permet de se concentrer sur l'évolution des résidus dans le temps, et ce, peu importe le "niveau" de ces résidus (voir figure 3.17 pour une illustration avec des individus fictifs ayant des niveaux de résidus différents, mais des trajectoires de résidus centrés similaires).

7. Une interaction avec le temps significative indique que la relation entre la variable prédictive et la mesure du construit étudié change au cours du temps (phénomène qualifié par les auteurs de repriorisation).

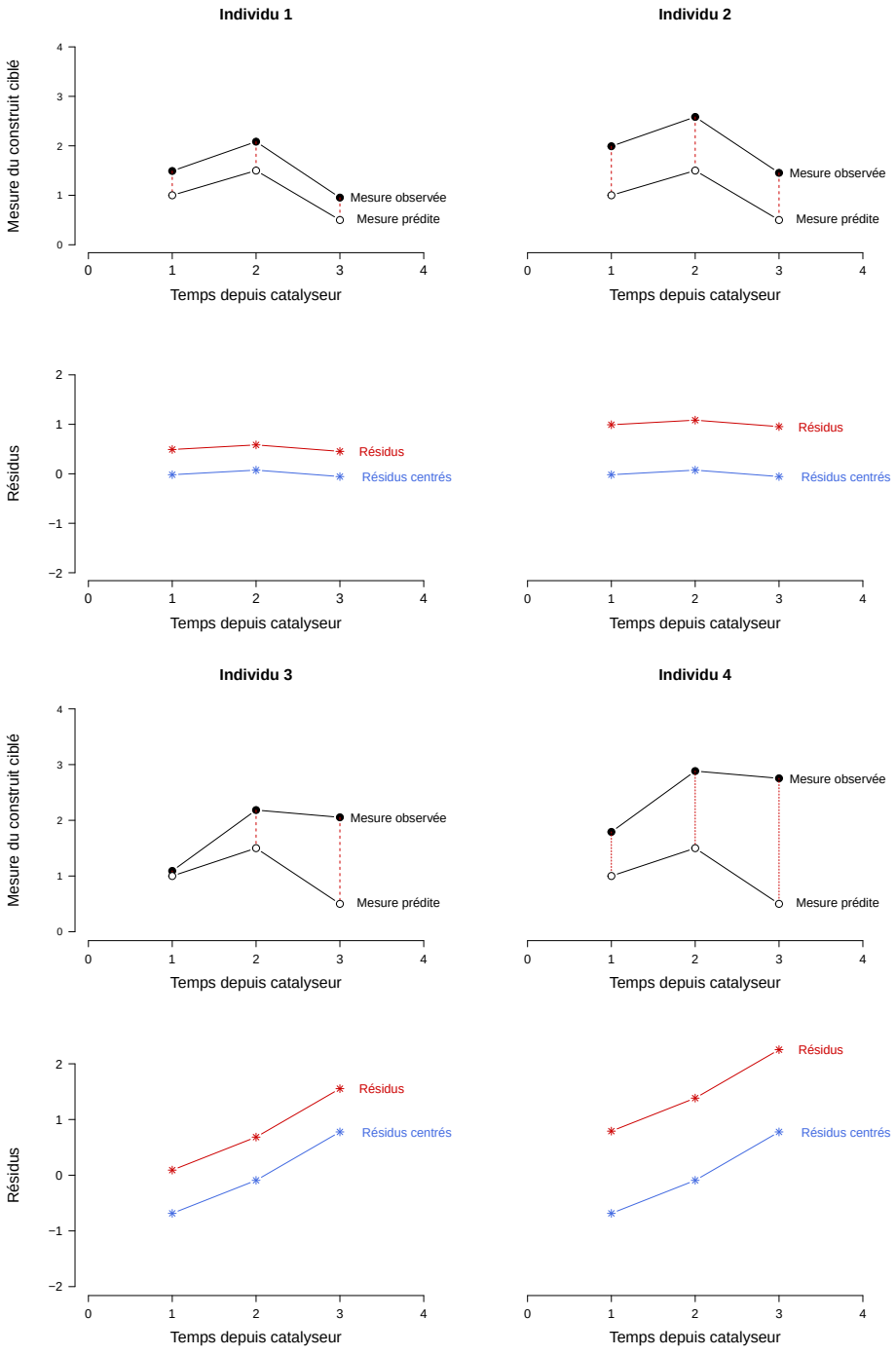


FIGURE 3.17 – Évolution des résidus (centrés) pour 4 individus fictifs

Note : Les individus 1 et 2 ont des mesures observées systématiquement supérieures aux valeurs prédites par le modèle (phénomène plus marqué pour l'individu 2). Comme les écarts entre les mesures observées et prédites restent consistants dans le temps, ces deux individus ont tous deux des résidus centrés qui fluctuent autour de 0.
Les individus 3 et 4 ont des mesures observées systématiquement supérieures aux valeurs prédites par le modèle, mais l'écart entre valeurs prédites et observées augmente avec le temps (phénomène plus marqué pour l'individu 4). Ces deux individus ont tous deux des résidus centrés qui augmentent.

Étape 3 : Étude des différentes trajectoires des résidus centrés

La dernière étape de la méthode consiste à identifier les différentes trajectoires de résidus centrés dans l'échantillon. Cette identification des différentes trajectoires est réalisée grâce l'estimation de modèles mixtes à classes latentes (*Growth Mixture Model*, GMM) où les résidus centrés sont modélisés en fonction du temps écoulé depuis l'événement catalyseur. Ces modèles supposent que la population étudiée est constituée d'un mélange de groupes (classes latentes) ayant leur propre trajectoire de résidus centrés. En pratique, plusieurs modèles sont estimés où différents nombres de classes et différents types de trajectoires (linéaire, quadratique, cubique) sont considérés [8]. Le meilleur modèle est ensuite sélectionné sur la base des critères d'information Bayésien [81] et d'Akaike [152].

Hypothèse de la méthode :

Le *response shift* pourrait se manifester par des fluctuations importantes dans les résidus centrés des individus.

L'étude de la trajectoire des résidus pourrait permettre d'étudier la variabilité interindividuelle du *response shift*. Elle pourrait également apporter des informations sur la temporalité du *response shift*, en permettant de prendre en compte plusieurs temps de mesure. Cette méthode a par ailleurs été couplée à la procédure de Oort [139] par Salmon *et al.* afin d'identifier les formes de *response shift* impliquées dans les différentes classes latentes identifiées chez des patientes atteintes d'un cancer du sein [153].

8. Pour pouvoir modéliser une trajectoire linéaire, il faut au minimum trois temps de mesure. Il en faut au moins quatre pour modéliser une courbe.

Autres méthodes

Cette section sur les méthodes de détection statistiques du *response shift* n'est pas exhaustive, on peut également citer d'autres méthodes comme : l'analyse de l'importance relative [154] et la détection du *response shift* à l'aide de CART (*Classification and regression tree*) [155] ou de forêts aléatoires [156]. Une méthode basée sur l'évolution du nombre d'erreurs de Guttman a également été proposée par Blanchin *et al.* [148]. Cette méthode a fait l'objet de développements au cours de cette thèse. Elle est donc décrite dans le chapitre présentant ces travaux (chapitre 4).

Partie II : Travaux personnels de thèse

Chapitre 4

Vers une détection plus individuelle du *response shift* avec les erreurs de Guttman

Sommaire

4.1 Motivations et objectifs

Comme rapporté par Sébille *et al.*, la plupart des méthodes statistiques pour la détection du *response shift* supposent que tous les individus, ou sous-groupes d'individus connus à l'avance, expérimentent le *response shift* de la même façon [23]. Cela signifie que ces méthodes supposent que le *response shift* affecte l'ensemble des répondants (ou sous-groupes de répondants) sur les mêmes domaines (ou items), et que la temporalité, la magnitude et la direction du *response shift* est la même pour tous. Ces méthodes omettent donc le caractère variable et individuel de ce phénomène, alors que la survenue du *response shift* est probablement influencée par les caractéristiques culturelles et cliniques des répondants [23]. La prise en compte de cette variabilité est capitale pour une meilleure compréhension du *response shift* et de ses déterminants [157]. Ainsi, la détection du *response shift* à un niveau plus individuel représente aujourd'hui un champ d'intérêt pour la recherche dans ce domaine [23, 157].

En 2016, Blanchin *et al.* [148] - mes encadrants de thèse - ont suggéré que la survenue de *response shift* pourrait interférer avec la cohérence des réponses des individus. Plus précisément, ils ont émis l'hypothèse que les individus expérimentant du *response shift* entre deux temps de mesure pourraient avoir des réponses moins cohérentes au second temps t_2, comparées à leurs réponses au premier temps t_1. Afin de mesurer le degré de cohérence des réponses des individus, ils ont eu recours aux erreurs de Guttman. Pour chaque temps de mesure et chaque individu, les erreurs de Guttman ont été dénombrées en comptant le nombre de fois où l'individu atteignait une modalité de réponse, alors qu'il n'avait pas réussi à atteindre une modalité de réponse plus facile. Les difficultés des modalités de réponse, nécessaires à ce comptage, ont été définies en utilisant les données recueillies sur l'ensemble de l'échantillon au premier temps de mesure.

À partir d'une application sur données réelles, ces auteurs avaient identifié deux groupes d'individus : un premier groupe présentant un nombre d'erreurs de Guttman approximativement constant au cours du temps, et un second groupe présentant une augmentation du nombre d'erreurs de Guttman. Ces deux groupes avaient alors été respectivement qualifiés d'"individus

peu susceptibles d'expérimenter du *response shift*" et d'"individus susceptibles d'expérimenter du *response shift*".

Cette méthode non paramétrique, basée sur les erreurs de Guttman, leur avait semblée prometteuse pour une détection individuelle du *response shift*, au niveau des items. En effet, elle avait donné des résultats concordants avec une autre méthode de détection (paramétrique cette fois-ci) : l'algorithme ROSALI. Plus précisément, l'algorithme ROSALI avait détecté du *response shift* dans le second groupe (individus identifiés comme étant susceptibles d'expérimenter du *response shift*) et n'en avait pas identifié dans le premier [148]. Cette observation laissait donc entendre une certaine concordance entre les résultats de ces deux méthodes. Il s'agissait néanmoins d'une application sur données réelles, où la "vérité" n'est pas connue. Plusieurs questions peuvent donc se poser :

1. Le *response shift* induit-il effectivement une augmentation du nombre d'erreurs de Guttman chez les individus qui l'expérimentent ?

2. En se basant sur l'évolution du nombre d'erreurs de Guttman, peut-on distinguer les patients expérimentant du *reponse shift* des autres patients n'en expérimentant pas ?

3. Mis à part le *response shift*, d'autres phénomènes peuvent-ils être à l'origine d'une augmentation du nombre d'erreurs de Guttman ?

Les travaux présentés dans ce chapitre ont pour objectifs de déterminer : (i) si la recalibration (une des formes de *response shift*) se manifeste effectivement par une augmentation du nombre d'erreurs de Guttman au cours du temps, et (ii) si cette différence est discriminante ou non (*i.e.*, permet-elle de distinguer les individus expérimentant du *response shift* des autres ?).

Ces travaux sont restreints à la détection de la recalibration pour deux raisons. Premièrement, nous souhaitions nous concentrer sur l'étude d'échelles unidimensionnelles, la reconceptualisation ne pouvait donc pas être considérée. De plus, des questionnements sur la signification, l'interprétation et la pertinence d'une repriorisation ayant lieu au niveau des items ont été récemment soulevés [146, 149].

4.2 Détection de la recalibration à l'aide des erreurs de Guttman

Afin de formaliser la méthode de détection proposée par Blanchin *et al.* [148], nous allons considérer un exemple fictif, où un échantillon d'individus répond à un questionnaire visant à mesurer la vitalité à deux temps de mesure t_1 et t_2. Le questionnaire est composé de 3 items polytomiques avec 4 modalités de réponse (0, 1, 2 et 3) donnés ci-dessous (ces items sont tirés de la dimension "vitalité" du SF-36 [158], librement adaptés pour l'exemple) :

Items

1. Vous sentez-vous dynamique ?	❒ Pas du tout (0)	❒ Un peu (1)	❒ Assez (2)	❒ Beaucoup (3)
2. Avez-vous de l'énergie ?	❒ Pas du tout (0)	❒ Un peu (1)	❒ Assez (2)	❒ Beaucoup (3)
3. Êtes-vous fatigué(e) ?	❒ Beaucoup (0)	❒ Assez (1)	❒ Un peu (2)	❒ Pas du tout (3)

Pour des raisons de simplicité, nous allons considérer que le niveau de la variable latente "vitalité" des individus étudiés ne change pas entre les deux temps de mesure.

Étape n°1 : Ordonner les modalités de réponse à l'aide des données recueillies au temps t_1

La première étape de la méthode consiste à ordonner les modalités de réponse supérieures à 0 de la plus facile à la plus difficile, en utilisant les données recueillies au sein de l'échantillon au temps t_1. La répartition des réponses est donnée dans la figure 4.1 :

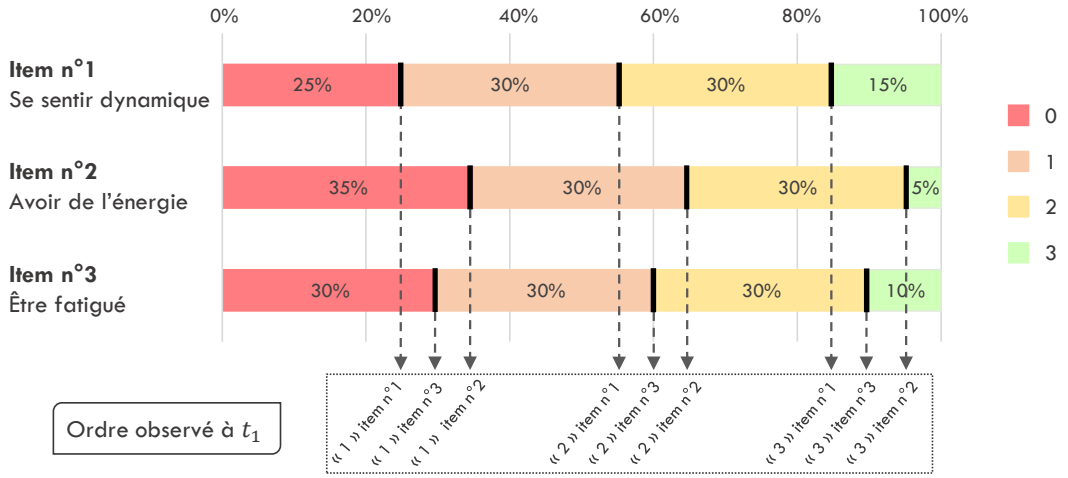


FIGURE 4.1 – Distribution des réponses de l'échantillon au temps t_1 (exemple fictif)

Une modalité de réponse est considérée comme facile si de nombreux individus l'ont atteinte. Elle est considérée difficile dans le cas contraire. En se basant sur la figure 4.1 (représentant la distribution des réponses de l'échantillon au temps t_1 pour l'exemple fictif) on remarque que la modalité la plus facile est la modalité "1" pour l'item n°1. En effet, 75% des individus l'ont atteinte (c.-à-d., ont répondu 1 ou plus pour l'item n°1). La deuxième modalité la plus facile est la modalité de réponse "1" de l'item n°3 (70% des individus ont répondu au moins "1" à cet item). La modalité suivante est la modalité "1" pour l'item n°2 (avec 65% de l'échantillon ayant choisi cette modalité ou une modalité supérieure). On continue ainsi de suite jusqu'à la modalité de réponse la plus "difficile" : la modalité "3" pour l'item n°2 (avec seulement 5% de l'échantillon qui réussit à atteindre cette modalité). Ces modalités de réponse ainsi ordonnées forment l'ordre de difficulté observé au premier temps de mesure. L'ordre complet est représenté dans le bas de la figure 4.1. L'exemple fictif considéré est ici volontairement simpliste.

Étape n°2 : Dénombrer les erreurs de Guttman aux deux temps de mesure t_1 et t_2 en considérant l'ordre observé à t_1

Cette deuxième étape consiste à compter le nombre d'erreurs de Guttman dans les réponses de chaque individu, aux deux temps de mesure t_1 et t_2, en se rapportant à l'ordre défini à l'étape précédente. Une erreur de Guttman survient dès qu'un patient atteint une modalité de réponse pour un certain item, alors qu'il n'arrive pas à atteindre une modalité de réponse plus facile provenant d'un autre item [159].

Pour illustrer cette définition, considérons un patient répondant au premier temps t_1 :
item n°1 : "0 = Pas du tout", item n°2 : "0 = Pas du tout" et item n°3 : "1 = Assez".

Ce patient a alors une erreur de Guttman à t_1. En effet, il a réussi à atteindre la modalité "1" pour l'item n°3, alors qu'il n'a pas réussi à atteindre la modalité "1" de l'item n°1 (cette dernière était pourtant plus facile). Cette erreur de Guttman est mise en évidence dans la figure 4.2.

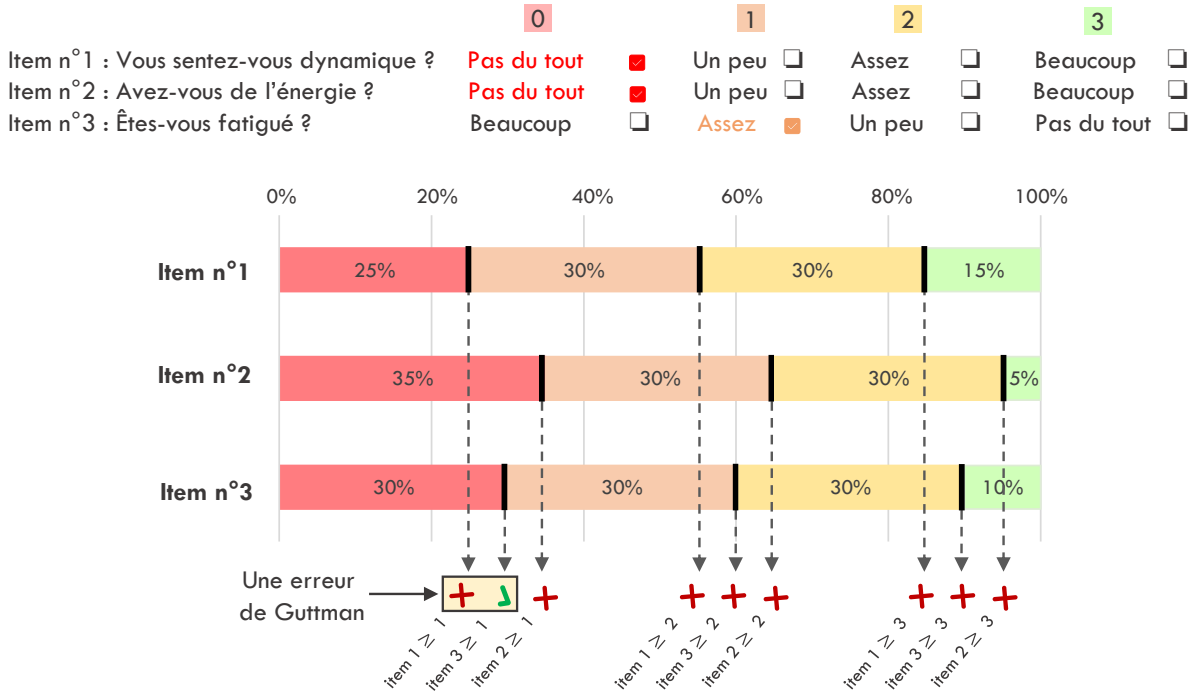


FIGURE 4.2 – Illustration de l'occurrence d'une erreur de Guttman (exemple fictif)

Notes : La coche verte dans le bas de la figure indique la modalité atteinte, tandis que les croix rouges indiquent les modalités non atteintes.

Dans la suite du manuscrit, on notera $EG^{(t)}_{ordre\ t_1}$ le nombre d'erreurs de Guttman au temps t ($t = t_1,\ t_2$), définies en utilisant l'ordre observé à t_1. Ces erreurs de Guttman permettent de mesurer les incohérences dans les réponses des patients, par rapport aux réponses observées sur l'ensemble de l'échantillon. On parle d'"erreurs" au sens de "déviations" par rapport au modèle déterministe de Guttman (qui spécifie que si une modalité de réponse est atteinte, alors toutes les modalités plus faciles doivent également l'être). Il n'y a bien évidemment pas d'"erreurs" à proprement parler dans les réponses des patients.

Étape n°3 : Déterminer pour chaque individu l'évolution du nombre d'erreurs de Guttman entre les deux temps (en utilisant l'ordre observé à t_1)

Pour cette dernière étape, on calcule pour chaque individu l'évolution du nombre erreurs de Guttman définie par :

$$I = EG^{(t_2)}_{ordre\ t_1} - EG^{(t_1)}_{ordre\ t_1}$$

Hypothèses de la méthode : Les individus n'expérimentant pas de recalibration devraient avoir un nombre d'erreurs de Guttman stable au cours du temps. En effet, leur perception du questionnaire n'ayant pas changé, les réponses qu'ils donnent au temps t_2 devraient toujours concorder avec l'ordre défini au départ. Ils devraient donc vérifier $I \approx 0$. Au contraire, les individus expérimentant de la recalibration pourraient voir leur nombre d'erreurs de Gutmann augmenter entre les deux temps : à cause de leurs changements de perception, les réponses qu'ils donnent au temps t_2 pourraient ne plus concorder avec l'ordre défini au départ. Ils vérifieraient donc $I > 0$.

Pour illustrer cette dernière hypothèse, considérons le patient introduit précédemment. Si ce patient a un niveau de variable latente stable entre les deux temps de mesure, mais qu'il s'est adapté à sa condition suite à de forts épisodes de fatigue survenus entre t_1 et t_2, il pourrait ainsi répondre à t_2 : item n°1 = "0", item n°2 = "0" et item n°3 = "2". Sa réponse à l'item n°3 "Êtes-vous fatigué(e) ?" n'est plus la même qu'au temps t_1 : il n'évalue plus de la même façon sa fatigue, sa métrique interne ayant changé. On comptabiliserait alors pour ce patient quatre erreurs de Guttman. Ainsi, pour cet individu, on observerait $I = 3 > 0$.

L'indicateur I présenté ci-dessus correspond à la transposition directe de la méthode proposée par Blanchin *et al.* [148]. Néanmoins, le nombre d'erreurs de Guttman présente une limite importante : il est directement impacté par le score de l'individu au questionnaire. En effet, un individu ayant un score extrême (proche de 0 ou proche du score maximal) aura un nombre maximal d'erreurs de Guttman potentiellement atteignable très restreint. Au contraire, un individu ayant un score intermédiaire aura un nombre maximal d'erreurs de Guttman potentiellement

atteignable plus important. Par exemple, en reprenant l'ordre observé à t_1 dans l'exemple fictif, les plages des valeurs possibles pour le nombre d'erreurs de Guttman sont données dans la figure 4.3 en fonction des scores des individus. Les profils de réponse correspondant aux *extremums* de ces plages sont explicités dans le tableau 4.1 pour chaque score. Il faut noter que ces résultats ne seraient pas les mêmes avec un autre ordre de difficulté. De plus, l'étendue des valeurs possibles pour le nombre d'erreurs de Guttman dépend également de la structure du questionnaire (nombre d'items et de modalités de réponse). En effet, plus il y a d'items et/ou de modalités de réponse, plus cette étendue s'agrandit.

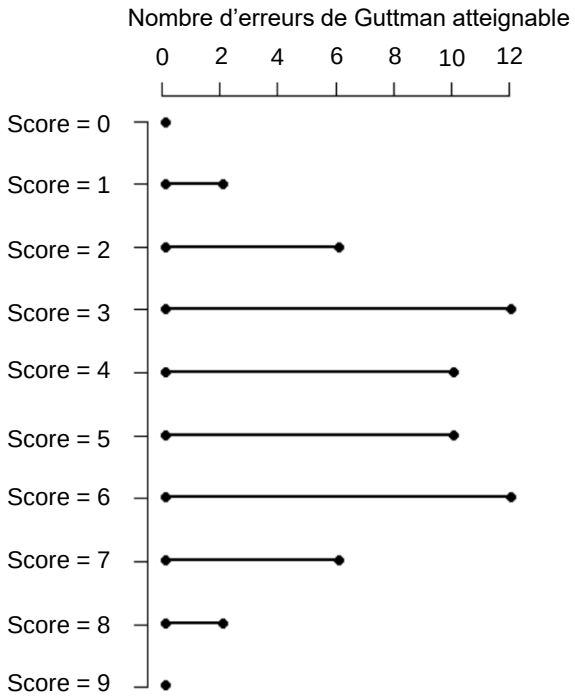


FIGURE 4.3 – Plages des valeurs possibles pour le nombre d'erreurs de Guttman en fonction de chaque score (exemple fictif)

TABLEAU 4.1 – Profils de réponse minimisant et maximisant le nombre d'erreurs de Guttman (exemple fictif)

	item 1 ≥ 1	item 3 ≥ 1	item 2 ≥ 1	item 1 ≥ 2	item 3 ≥ 2	item 2 ≥ 2	item 1 ≥ 3	item 3 ≥ 3	item 2 ≥ 3	Nb EG
(a) Profils de réponse n'engendrant pas d'erreurs de Guttman										
Score = 0	✗	✗	✗	✗	✗	✗	✗	✗	✗	0
Score = 1	✓	✗	✗	✗	✗	✗	✗	✗	✗	0
Score = 2	✓	✓	✗	✗	✗	✗	✗	✗	✗	0
Score = 3	✓	✓	✓	✗	✗	✗	✗	✗	✗	0
Score = 4	✓	✓	✓	✓	✗	✗	✗	✗	✗	0
Score = 5	✓	✓	✓	✓	✓	✗	✗	✗	✗	0
Score = 6	✓	✓	✓	✓	✓	✓	✗	✗	✗	0
Score = 7	✓	✓	✓	✓	✓	✓	✓	✗	✗	0
Score = 8	✓	✓	✓	✓	✓	✓	✓	✓	✗	0
Score = 9	✓	✓	✓	✓	✓	✓	✓	✓	✓	0
(b) Profils de réponse engendrant le maximum d'erreurs de Guttman										
Score = 0	✗	✗	✗	✗	✗	✗	✗	✗	✗	0
Score = 1	✗	✗	✓	✗	✗	✗	✗	✗	✗	2
Score = 2	✗	✗	✓	✗	✗	✓	✗	✗	✗	6
Score = 3	✗	✗	✓	✗	✗	✓	✗	✗	✓	12
Score = 4	✗	✓	✓	✗	✗	✓	✗	✗	✓	10
Score = 5	✗	✓	✓	✗	✓	✓	✗	✗	✓	10
Score = 6	✗	✓	✓	✗	✓	✓	✗	✓	✓	12
Score = 7	✓	✓	✓	✗	✓	✓	✗	✓	✓	6
Score = 8	✓	✓	✓	✓	✓	✓	✗	✓	✓	2
Score = 9	✓	✓	✓	✓	✓	✓	✓	✓	✓	0

Notes :

- Nb EG : Nombre d'erreurs de Guttman ;

– Une croix ✗ dans la colonne item $j \geq k$ indique que l'individu a choisi une modalité de réponse inférieure à k pour l'item j (il n'a pas réussi à atteindre la modalité k) ;

– Une coche ✓ dans la colonne item $j \geq k$ indique que l'individu a choisi la modalité de réponse k ou une moalité supérieure pour l'item j (il a réussi à atteindre la modalité k) ;

– Pour déterminer le nombre d'erreurs de Guttman associé à chaque profil de réponse, il suffit de compter le nombre de fois où chaque symbole ✓ est précédé par un symbole ✗, puis faire la somme des valeurs obtenues.

Par exemple, dans le profil de réponse maximisant le nombre d'erreurs de Guttman pour un score égal à 2 : le premier succès ✓ est précédé par deux échecs ✗ ✗ et le second succès ✓ est précédé par 4 échecs ✗ ✗ ✗ ✗. Ainsi, il y a 2 + 4 = 6 erreurs de Gutttman dans ce profil de réponse.

Pour pallier ces deux limites, il est possible de normer le nombre d'erreurs de Guttman : pour chaque individu et à chaque temps, le nombre d'erreurs de Guttman normé correspond au nombre d'erreurs de Guttman observé divisé par le nombre maximal d'erreurs de Guttman qu'il était possible d'observer étant donné le score de l'individu [1] :

$$EGnorm^{(t)}_{ordre\ t_1} = \frac{EG^{(t)}_{ordre\ t_1}}{max\left(EG^{(t)}_{ordre\ t_1} \mid Score^{(t)}\right)}$$

Un second indicateur I_{norm}, défini ci-dessous, a donc été étudié en parallèle de l'indicateur I :

$$I_{norm} = EGnorm^{(t_2)}_{ordre\ t_1} - EGnorm^{(t_1)}_{ordre\ t_1}$$

4.3 1re étude de simulation

Nous avons mis en place une étude de simulation de type "preuve de concept" pour confirmer l'hypothèse de la méthode précédemment présentée et explorer les capacités discriminantes des indicateurs I et I_{norm}.

Les études de simulation stochastiques (ou de Monte-Carlo) jouent un rôle clé dans la recherche en statistique, car elles permettent d'obtenir des données empiriques sur la pertinence, la validité, les performances ou encore la robustesse d'une ou plusieurs méthodes d'analyse [161]. En fonction de ces données, des recommandations peuvent alors être émises quant à l'utilisation de chacune des méthodes d'analyse évaluées (Dans quel(s) cas de figure une méthode d'analyse présente de bonnes ou de mauvaises performances ? Y a-t-il une méthode à privilégier ?). Le principe d'une étude de simulation réside dans la création de données fictives où l'on connaît la

1. Le nombre maximal d'erreurs de Guttman qu'il est possible d'atteindre s'obtient facilement avec des items binaires. En revanche, il ne se détermine pas aussi directement avec des items polytomiques. Pour le déterminer, il est possible d'utiliser l'algorithme proposé par Emons [160]. Au cours de cette thèse, un module Stata a été programmé afin de pouvoir calculer automatiquement le nombre d'erreurs de Guttman, le nombre maximal d'erreurs de Guttman et le nombre d'erreurs de Guttman normé à partir d'un tableau de données.

"vérité" et où l'on maîtrise les valeurs des paramètres. Par exemple, dans le cas de la détection du *response shift* à un niveau individuel, on peut générer des données où l'on connaît et l'on contrôle par avance : la taille de l'échantillon, la structure du questionnaire, les paramètres liés à la variable latente, les individus expérimentant (ou non) de la recalibration, les items sur lesquels cette recalibration se manifeste, la forme de la recalibration ou encore sa taille. En faisant varier les valeurs des paramètres et les cas de figure considérés, il est possible d'étudier les propriétés de la/des méthode(s) à l'étude dans des situations variées, appelées scénarios. Afin de prendre en compte la fluctuation d'échantillonnage, chaque scénario est répliqué un certain nombre de fois. Les jeux de données correspondant à ces réplications sont ensuite analysés avec la/les méthode(s) étudiée(s). On détermine alors si ce qui a été simulé a été retrouvé. C'est là le point clé des études de simulation : contrairement aux données réelles, on sait à l'avance ce que la méthode est censée retrouver.

Dans notre étude de simulation, l'objectif était de déterminer si la survenue de recalibration s'accompagne effectivement d'une augmentation du nombre d'erreurs de Guttman dans les réponses des patients qui l'expérimentent. On s'est également intéressé à la capacité discriminante de nos indicateurs. En d'autres termes, on a cherché à déterminer si ces indicateurs permettaient de distinguer les individus pour qui on a simulé de la recalibration des autres (pour qui on n'en a pas simulé). Cette étude de simulation a fait l'objet d'une publication dans le journal *Quality of Life Research* (section spéciale "NIRT : *non-parametric item response theory*") présentée dans la section suivante [162].

4.3.1 Article : Evaluation of the link between Guttman errors and response shift at the individuel level

Terminologie du manuscrit : Dans tout le manuscrit, les erreurs de Guttman sont désignées par l'abréviation GE (pour Guttman errors) et non pas par l'abréviation EG. La terminologie "patients affected by recalibration" a été utilisée pour désigner les individus expérimentant de la recalibration (pour l'étude de simulation, il s'agit des individus pour lesquels on a simulé de

la recalibration). La terminologie "items affected by recalibration" a été utilisée pour désigner les items sur lesquels se manifeste la recalibration (pour l'étude de simulation, il s'agit des items pour lesquels on a simulé de la recalibration, c'est-à-dire les items dont on a fait varier les paramètres de seuil entre les deux temps de mesure).

Quality of Life Research
https://doi.org/10.1007/s11136-021-03015-9

SPECIAL SECTION: NON-PARAMETRIC IRT

Evaluation of the link between the Guttman errors and response shift at the individual level

Yseulys Dubuy[1] · Véronique Sébille[1] · Marie Grall-Bronnec[1,2] · Gaëlle Challet-Bouju[1,2] · Myriam Blanchin[1] · Jean-Benoit Hardouin[1]

Accepted: 4 October 2021

Abstract

Purpose Methods for response shift (RS) detection at the individual level could be of great interest when analyzing changes in patient-reported outcome data. Guttman errors (GEs), which measure discrepancies in respondents' answers compared to the average sample responses, might be useful for detecting RS at the individual level between two time points, as RS may induce an increase in the number of discrepancies over time. This study aims to establish the link between recalibration RS and the change in the number of GEs over time (denoted index I) via simulations and explores the discriminating ability of this index.

Methods We simulated the responses of individuals affected or not affected by recalibration RS (defined as changes in the patients' standard of measurement) to determine whether simulated individuals with recalibration had a greater change in the number of GEs over time than individuals without recalibration. The effects of factors related to the sample, the questionnaire structure and recalibration were investigated. As an illustrative example, the change in the number of GEs was computed in patients suffering from eating disorders.

Results Within simulations, simulated individuals affected by recalibration had, on average, a greater change in the number of GEs over time than did individuals without RS. Some of the parameters related to the questionnaire structure and recalibration magnitude appeared to have substantial effects on the values of I. Discriminating abilities appeared, however, globally low.

Conclusion Some evidence of the link between recalibration and the change in GEs was found in this study. GEs could be a valuable nonparametric tool for RS detection at a more individual level, but further investigation is needed.

Keywords Response shift · Guttman errors · Recalibration · Individual level

Introduction

Patient-reported outcomes (PROs) are increasingly being used in longitudinal studies to take into account patients' perspectives on healthcare and to assess perceived health changes over time [1]. PROs are often investigated via questionnaires (directly completed by patients), including several items usually grouped into domains (e.g., physical, emotional, social functioning, etc.). The unobservable attributes targeted by these questionnaires (such as fatigue and anxiety) are assumed to be represented by nonobservable continuous variables known as "latent variables".

It is usually assumed that patients' perception of the concept of interest, the questions, and the response categories remain the same over time and that observed changes reflect changes in the latent variable (i.e., longitudinal measurement invariance). Hence, patients' responses at two different times are assumed to be directly comparable. However, the cognitive [2] and affective processes involved in questionnaires completion are complex, and PRO changes in longitudinal studies can be difficult to analyze and interpret. Moreover, the assumption of invariance may be questionable, particularly in the context of chronic diseases where patients regularly adapt to their life circumstances. Hence, there might be changes in the meaning of patients' self-evaluations of a target construct, referred to as response shift (RS) [3]. RS is usually assumed to have 3 manifestations: (1) recalibration

✉ Yseulys Dubuy
Yseulys.Dubuy@univ-nantes.fr

[1] INSERM U1246, SPHERE University of Nantes, University of Tours, Nantes, France

[2] Addictive Medicine and Psychiatry Department, CHU Nantes, Nantes, France

Published online: 17 October 2021

(changes in the patient's internal standards of measurement), (2) reprioritization (changes in the relative importance a patient gives to a certain component of the target construct, e.g., social functioning, which can become more important than physical functioning) and (3) reconceptualization (changes in the patient's definition of what is being measured). It is essential to assess changes experienced by patients taking into account RS to avoid measurement bias[1] and to detect and quantify RS in a reliable and unbiased manner because of its possible association with patients' adaptation [3–5].

Several statistical methods have been proposed for RS detection. Until recently, these methods were all developed and applied at the domain level, which means that analyses are performed on the domain scores of a multidimensional scale. The most widely used method is Oort's procedure based on structural equation modeling (SEM) [6], which allows for the detection of the three manifestations of RS.

Recently, interest in exploring RS at the item level has increased [7]. Indeed, item-level methods could provide an interesting and complementary perspective when investigating RS, as domain-level analyses may sometimes not appropriately reflect what is occurring at the item level, especially if RS has opposite effects on different items. Among the item-level methods, ROSALI (RespOnse Shift ALgorithm at Item-level) based either on Item Response Theory (IRT) or Rasch Measurement Theory (RMT) has been proposed [8, 9]. IRT-based ROSALI aims at detecting RS between two measurement occasions by allowing the item parameters of a longitudinal generalized partial credit model to vary over time (i.e., item discrimination parameter and item difficulty parameter). Changes in discrimination parameters and difficulty parameters are assumed to be indicative of reprioritization and recalibration RS, respectively. RMT-based ROSALI follows the same algorithm as IRT-based ROSALI but relies on a longitudinal partial credit model; hence, it enables the detection of recalibration only. For SEM, Oort's procedure has also been applied at the item level in different ways to detect recalibration and reprioritization [10–14].

Most methods, such as Oort's procedure and ROSALI, assume homogeneous RS within the sample or subgroups of patients known in advance. This assumption is probably too restrictive. Indeed, RS is likely to occur at different times and have different manifestations and various effect sizes among patients. In addition, whether at the domain or item level, these methods are parametric and thus rely on assumptions (e.g., normal latent variable distribution, normally distributed item responses, and link functions) that might be too restrictive.

A method relaxing these assumptions and focusing on the item and individual levels could be of great practical value. At the item level, Blanchin et al. suggested that RS (by interfering with patients' internal standards of measurement and life priorities) might induce discrepancies in individuals' responses over time relative to sample responses [15]. Based on this assumption, they identified, in a real data application, two groups of patients: one with an approximately constant number of discrepancies over time (assumed unlikely to be affected by RS) and another with an increasing number of discrepancies (assumed likely to be affected by RS). The discrepancies were measured nonparametrically using the (weighted) Guttman errors (GEs), which were obtained by comparing the individual responses to the distribution of the sample responses. The ROSALI algorithm detected RS in the subgroup of patients identified as likely to be affected by RS and did not detect RS in the other subgroup. Hence, it was hypothesized that GE could be a useful nonparametric tool for item-level RS detection at a more individual level; however, this was an empirical example, and one may wonder whether RS actually leads to an increase in the number of GE and to what extent RS has occurred if the number of GE increases over time. In addition, no information is currently available on the ability of the change in the number of GEs to discriminate patients affected by RS from the others.

We performed an exploratory simulation study aiming at (1) establishing the link between recalibration RS and the change in the number of GEs over time (i.e., determining whether recalibration comes with an increase in the number of GEs over time) in various scenarios representative of clinical research studies, (2) determining the simulation parameters associated with this change, and (3) providing some data on the discriminating ability of this GEs-based index in the case of recalibration RS.

Methods

We chose to ascribe our GEs-based index to the detection of recalibration in this first simulation study for the following reasons: (1) we wanted to focus on unidimensional scales; hence, reconceptualization could not be considered for now (2) the potential different meanings and interpretations of reprioritization at the item level from a methodological or conceptual perspective have already been raised [7], and Blanchin et al. went even further by questioning the pertinence of the concept of reprioritization at the item level [9].

Guttman errors for recalibration RS detection

Let us consider a fictitious sample of individuals responding to a scale composed of 4 items, with 4 response categories, namely, 0, 1, 2 and 3, at two measurement occasions

[1] i.e., nonrandom errors in the latent variable estimates.

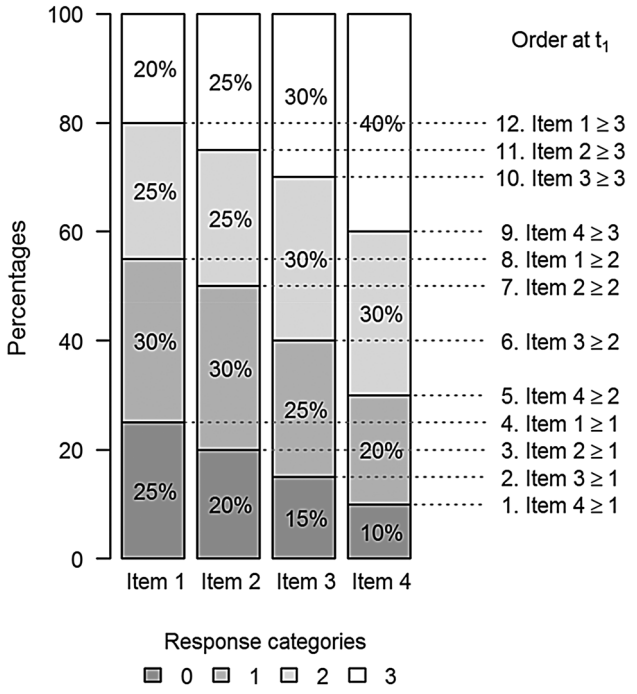


Fig. 1 Percent stacked barchart showing the proportion of patients choosing each response category (0, 1, 2, 3) for each item (item 1, item 2, item 3, item 4) at t_1 (fictitious example)

(t_1 and t_2). Let us assume that the following assumptions hold at both time points: (1) unidimensionality (i.e., all items in the questionnaire measure the same latent variable), (2) local independence (i.e., given the latent variable, item responses are independent), and (3) monotonicity (i.e., when the latent variable increases, the probability of obtaining at least score x on item j does not decrease). For simplicity, let us also assume that the underlying latent variable remains stable over time for all patients. The distributions of the sample responses at t_1 are given in Fig. 1.

Guttman errors

To define GEs, we have to order all the response categories above 0 based on their difficulty. The term "difficulty" refers to how frequently a response category is endorsed: the less a category is endorsed, the more difficult it is considered to be. At t_1, all response categories greater than 0 can be ordered from the easiest to the most difficult. The difficulty of the response category x ($= 1, 2, 3$) from item j ($= 1, 2, 3, 4$) at t_1 is the proportion of patients scoring below x for item j at t_1.

In our example, we can observe that the easiest positive response category is "1" of item 4 since only 10% of the sample scored below this option. The second easiest category is "1" of item 3 (15% of the sample scored below this option). And so on, until the most difficult category which is "3" of item 1, since 80% of the sample scored below this option (i.e., 0, 1 or 2). Null categories are not included in this order since all patients endorse them. The order so defined is called the difficulty order observed at t_1. The term "difficulty" is sometimes referred to as "popularity" [16], easy and difficult response categories are then called "popular" and "unpopular", respectively.

GEs measure discrepancies in individual patient responses compared to the distribution of the sample responses. A GE occurs every time a patient endorses a response category for a given item, while he/she does not endorse an easier response category (for another item) [17]. For instance, a patient who responded item $1 = 2$, item $2 = 1$, item $3 = 2$ and item $4 = 2$ at t_1 has one GE according to the order defined at t_1. Indeed, he/she endorsed category "2" for item 1 but not category "2" for item 2. A formal definition of GE can be found in Emons' works [18].

Guttman errors and recalibration RS

Let us now assume that recalibration is observed on item 4 for half the sample between t_1 and t_2 (same manifestation and effect size for all affected patients) and that its three positive response categories' difficulties have increased at t_2.

At time t_2, patients without RS should give similar responses to those at t_1 since neither their latent variable nor their perception of the response categories have changed. Hence, among patients without RS, the GEs according to the order at t_1 is expected to remain the same over time. In contrast, at time t_2, responses for item 4 from patients affected by recalibration should deviate from the distribution of the sample responses observed at time t_1. The order observed at t_1 should no longer fit their responses due to recalibration, inducing more discrepancies. Hence, the number of GE (based on the order at t_1) of patients affected by recalibration is expected to increase over time. For example, if the patient previously introduced was affected by the recalibration on item 4, he/she could have responded at the second time point: item $1 = 2$, item $2 = 1$, item $3 = 2$ and item $4 = 0$ (due to recalibration, the response categories of item 4 became more difficult to endorse). According to the order defined at t_1, he/she has eight GEs at t_2 (instead of one at t_1). His/her overall sum score (computed by summing item responses) has changed but reflects the occurrence of recalibration and not a latent variable change.

Hence, counting the number of GEs over time using the order observed at t_1 could help identify recalibration. Indeed, patients without recalibration should have an approximately constant number of GEs over time, while an increase should be observed among patients with recalibration. We introduced a GEs-based index: the change over time in the number of GEs computed using the difficulty order defined at t_1 (denoted I), defined as follows:

Table 1 Simulation parameters

Sample and questionnaire	
Sample size (N)	$N = 100;200;300$
Proportion of patients affected by recalibration (p)	$p = 0.25;0.5;0.75$
Number of items (J)	$J = 4;7$
Number of response categories/item (M)	$M = 4;7;10$
Latent variable (Θ)	
Mean at time t_1 (μ_1)	$\mu_1 = 0$
Mean change ($\Delta = \mu_2 - \mu_1$)	$\Delta = -0.2;0;0.2$
Variance (σ_1^2, σ_2^2)	$\sigma_1^2 = \sigma_2^2 = 1$
Covariance between the two measurement occasions ($\sigma_{1,2}$)	$\sigma_{1,2} = 0.6$
Recalibration response shift size (η)	
UR	$\eta = -1$
NUR	$\eta = 1.8$
Items selected to show recalibration	
$J = 4$	
1 item affected	Item 3
2 items affected	Items 3 and 4
$J = 7$	
1 item affected	Item 5
2 items affected	Items 6 and 7
3 items affected	Items 4, 6 and 7

$$I = number\ of\ GEs^{(t_2)}_{order\ t_1} - number\ of\ GEs^{(t_1)}_{order\ t_1}$$

where $number\ of\ GEs^{(t^*)}_{order\ t_1}$ denotes the number of GEs observed at t^* using the difficulty order defined at t_1. It should be noted that at the second time point, the Guttman errors are computed using the ordering defined at t_1, hence there are not Guttman errors in the conventional sense, but slight adaptations of Guttman errors. The index I can be computed for each individual. We can expect that $I \approx 0$ among patients without RS and $I \geq 0$ among those with recalibration.

The link between recalibration RS and GEs was explored using a simulation study. Different parameters commonly encountered in analyses of PRO data were explored to determine their effect on the values of I. In a second step, we assessed the ability of I to discriminate patients affected by recalibration from others.

Simulation study

Data simulation

We simulated the responses of N individuals to a unidimensional questionnaire composed of J polytomous items with M response categories, numbered from 0 to $M - 1$, at two different time points. Endorsing difficult response categories was assumed to be manifestations of a high latent variable level. The longitudinal partial credit model (LPCM) [19] was chosen to generate data since it allowed modelling response category probabilities of polytomous items forming a unidimensional scale across time and provided a possibility to simulate recalibration for a changing proportion of individuals. All simulation parameters chosen for data generation are given in Table 1. Additional information on the simulation implementation is provided in Appendix 1.

Recalibration operationalization

Recalibration was operationalized as changes over time in the LPCM difficulty parameters [8, 9]. Recalibration may be uniform (UR: a change in all difficulty parameters of a given item in the same direction and to the same extent) or nonuniform (NUR: changes occur in various directions and intensities).

At the second time point t_2, recalibration was simulated as follows:

- Only one type of recalibration (UR or NUR) with the same size per data set was considered.
- The proportion p of the sample that was affected by recalibration was variable.
- The items affected by this recalibration were randomly selected (the same for all individuals affected by recalibration).

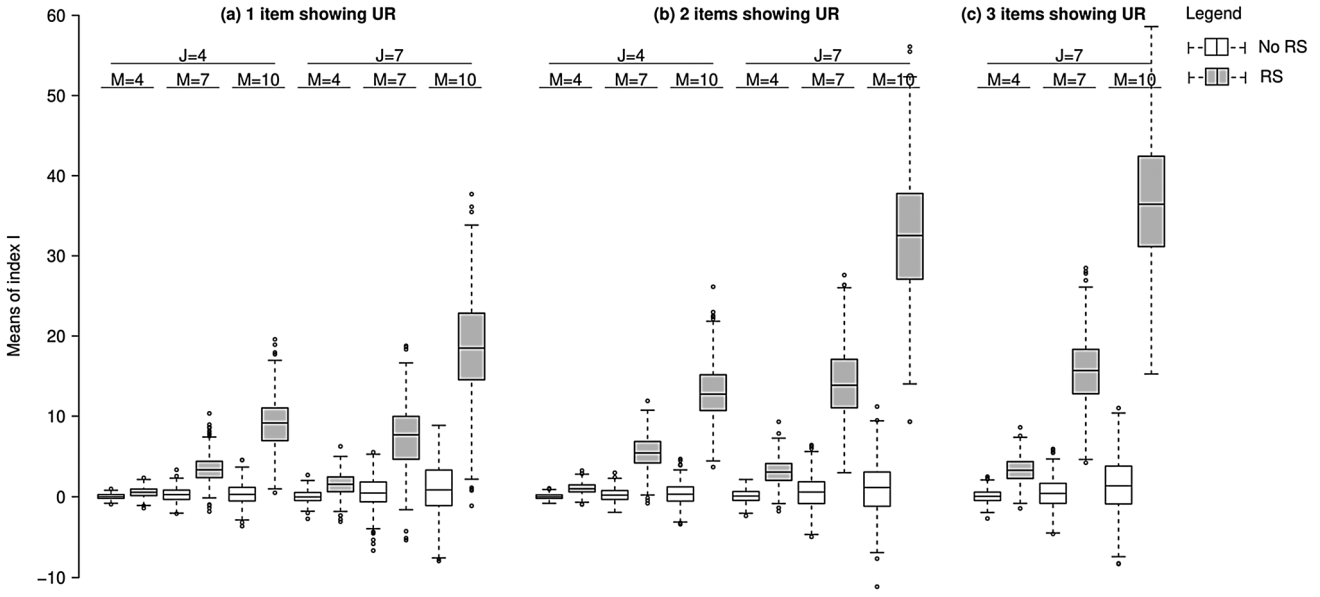


Fig. 2 Boxplots of the 500 mean values of index *I* obtained for each scenario among simulated patients affected by response shift (in white) and among simulated patients not affected by response shift (in grey). Each pair of boxplots corresponds to one scenario. Subset of scenarios considered: $N = 200$ (sample size), $p = 25\%$ (proportion of patients affected by response shift), $\Delta = -0.2$ (average change in the latent variable over time); uniform recalibration (UR). *RS* response shift, *J* number of items, *M* number of response categories per item

To generate the responses of the simulated patients affected by UR at t_2, all the difficulty parameters of the item(s) affected by recalibration decreased (–1), making the associated response categories easier. For simulated patients affected by NUR, difficulty parameters were differentially shifted at t_2 by values ranging from 0 to 2η, with $\eta = 1.8$. The first positive response category kept the same difficulty parameter over time, while other categories became more difficult. For simulated patients not affected by RS, the difficulty parameters remained constant over time.

We aimed to investigate the effect of recalibration RS-related factors (such as the number of items with recalibration J_{RS}, the proportion of the sample that was affected by recalibration p, and the recalibration type: UR and NUR) but also more global simulation parameters (the sample size N, the number of items in the questionnaire J, the number of response categories per item M, and the average change in the latent variable over time Δ). Simulation parameters were chosen to be representative of clinical research studies (Table 1).

The combination of all the simulation parameter values led to a total of 810 scenarios, and each of them was replicated 500 times.

Statistical analysis

Within the 500 replications of each scenario, index *I* was computed for each individual. For each scenario, the boxplots of the 500 mean values of index *I* obtained respectively among simulated patients affected and not affected by recalibration RS were plotted.

Over all replications, the discriminating abilities associated with the change in the number of GEs over time (i.e., index *I*) were estimated by the area under the receiver operating characteristic curve (AUROC), where response shift was the response variable and index *I* was the explanatory variable. Boxplots of the 500 AUROCs were also plotted (one per scenario).

Stata software release 15 was used for data generation (*simirt* module) and statistical analyses (StataCorp, 2015). Graphics were realized using the 3.5.3 version of R software (R Core Team, 2019).

Results

Simulation study

A small subset of scenarios is selected to present a representative portrayal of the variability of index *I* across experimental conditions ($N = 200$, 25% of the sample affected by recalibration, negative average change in the latent variable over time $\Delta = -0.2$; these values were chosen to approach the empirical example). Of note, all results are available in Online Resource 1.

Boxplots of the 500 average values of *I* among simulated patients with/without recalibration obtained for every scenario where $N = 200$, 25% of the sample was affected by

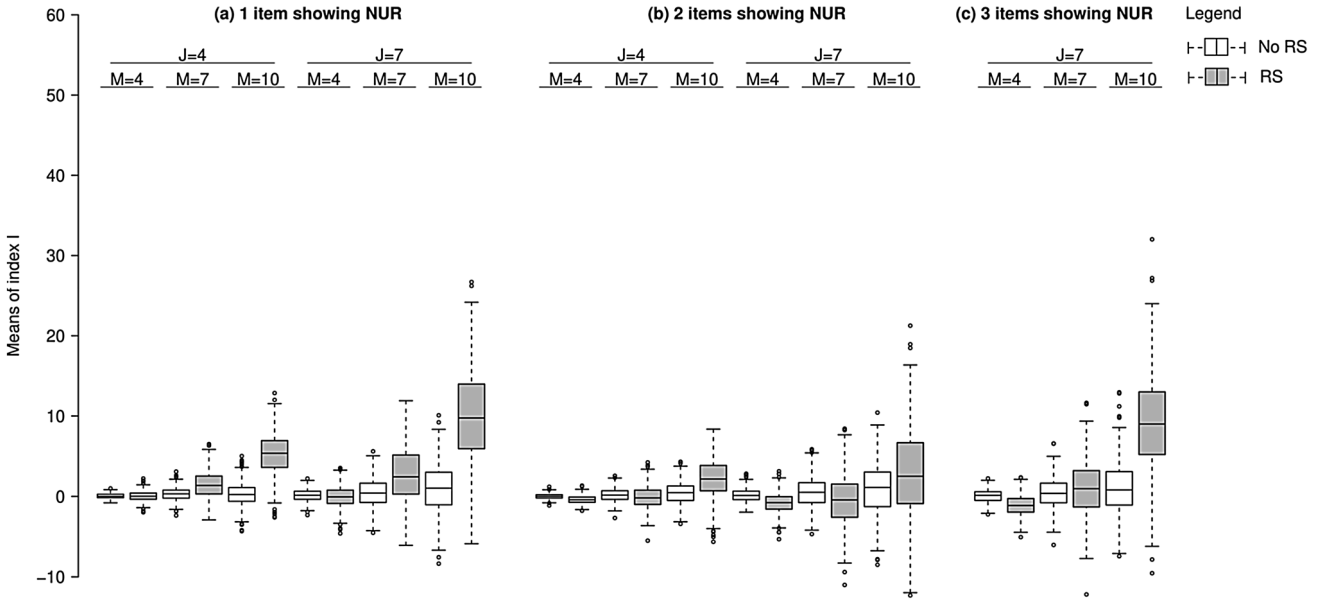


Fig. 3 Boxplots of the 500 mean values of index I obtained for each scenario among simulated patients affected by response shift (in white) and among simulated patients not affected by response shift (in grey). Each pair of boxplots corresponds to one scenario. Subset of scenarios considered: $N = 200$ (sample size), $p = 25\%$ (proportion of patients affected by response shift), $\Delta = -0.2$ (average change in the latent variable over time); nonuniform recalibration (NUR). *RS* response shift, J number of items, M number of response categories per item

recalibration RS, $\Delta = -0.2$ (negative average change in the latent variable over time) and RS = uniform recalibration are given in Fig. 2 according to the number of items affected by recalibration (J_{RS}), the number of items (J) and the number of response categories/item (M). Graphs under the same simulation conditions but for scenarios with nonuniform recalibration are given in Fig. 3.

Among simulated patients not affected by recalibration, the means of I fluctuated around values close to 0, regardless of the scenario considered. A slight increase in the means of I could, however, be observed when J and M increased. Among simulated patients with UR, the means of I fluctuated around positive values. These values remained low for scenarios with $M = 4$; however, they rose sharply when M increased. This rise was larger when J and J_{RS} were large. For simulated patients affected by NUR ($\eta = 1.8$), similar effects as those observed among simulated individuals affected by UR were observed for the average index values, but the trends were much less pronounced.

For each scenario, the dispersion of the means of I increased with J and M; the larger the overall number of response categories was, the wider the range of possible values for the number of GEs. This dispersion decreased logically as N increased.

Boxplots of the 500 AUROCs obtained for every scenario where $N = 200$, 25% of the sample was affected by recalibration RS, and $\Delta = -0.2$ (negative average change in the latent variable over time) are given in Fig. 4 (for uniform recalibration) and Fig. 5 (for nonuniform recalibration) according to the number of items affected by recalibration (J_{RS}), the number of items (J) and the number of response categories/item (M).

For UR, the discriminating abilities of I appeared to be low over all scenarios, particularly when $M = 4$. Indeed, in these cases, the average AUROC remained under 0.60. A slight increase could nonetheless be observed with increasing M and J_{RS}. For instance, the scenario with 3 items affected by recalibration, $J = 7$ and $M = 10$ resulted in an average AUROC close to 0.70. For NUR, the same phenomena were observed, but the AUROC values were even lower.

Illustrative example

To illustrate these results, we used a longitudinal study called EVALADD, which takes place at the Addictive Medicine and Psychiatry Department of Nantes University Hospital (France). The EVALADD cohort follows patients starting treatment for a behavioral addiction in order to assess the determinants of addictive disorders and, consequently, to improve therapies and preventive strategies. For this analysis, we focused on patients suffering from eating disorders (EDs) included between September 2012 and October 2016 (ED diagnoses were established according to the DSM-IV criteria [20] and explored via the French version of the Mini International Neuropsychiatric Interview [21, 22]). Patients completed self-reported questionnaires,

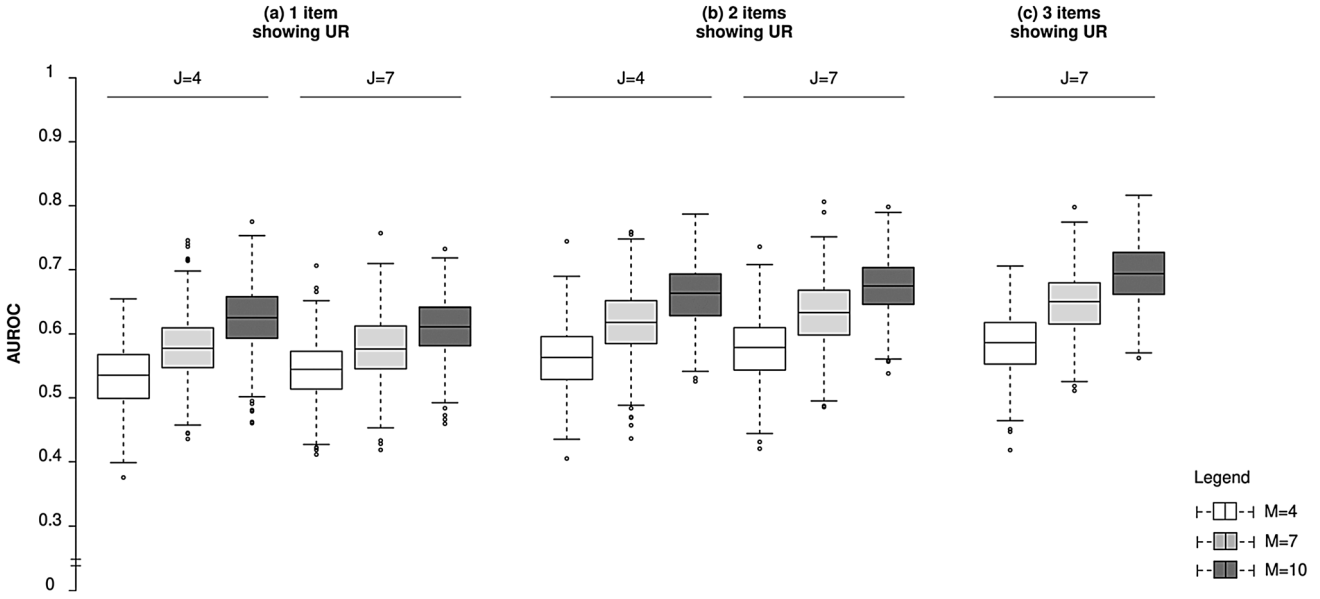


Fig. 4 Boxplots of the 500 AUROCs associated with I obtained for each scenario. Subset of scenarios considered: $N = 200$ (sample size), $p = 25\%$ (proportion of patients affected by response shift), $\Delta = -0.2$ (average change in the latent variable over time), uniform recalibration (UR). *AUROC* area under the receiver operating characteristic curve, J number of items, M number of response categories per item

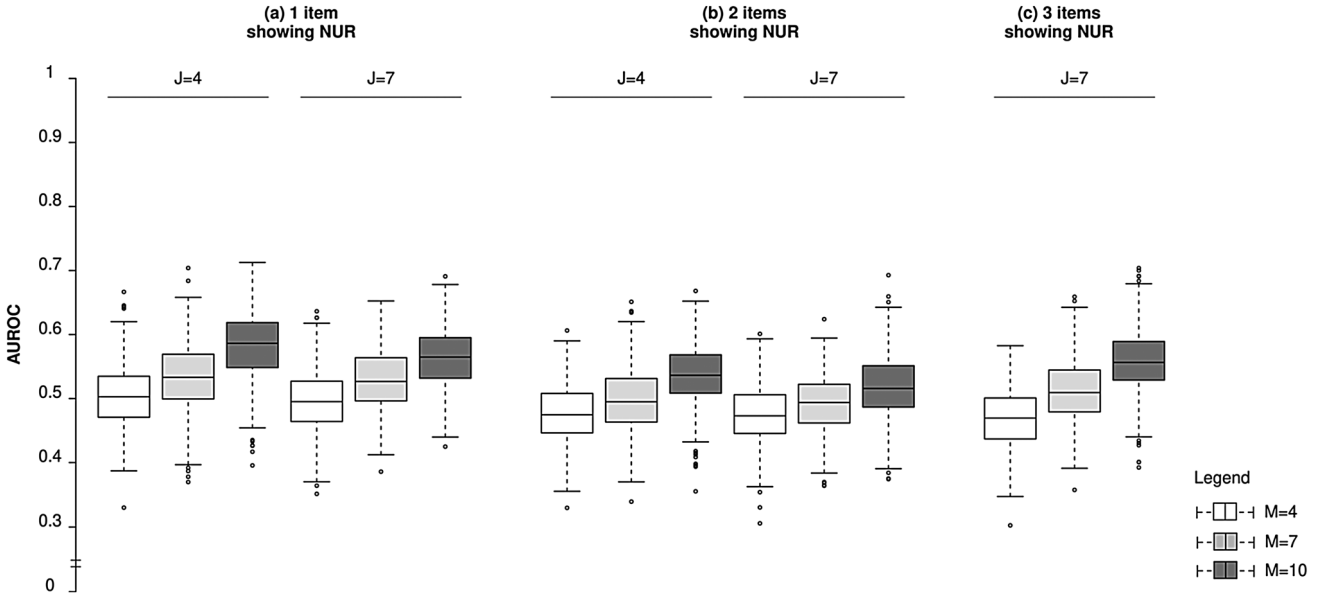


Fig. 5 Boxplots of the 500 AUROCs associated with I obtained for each scenario. Subset of scenarios considered: $N = 200$ (sample size), $p = 25\%$ (proportion of patients affected by response shift), $\Delta = -0.2$ (average change in the latent variable over time), nonuniform recalibration (NUR). *AUROC* area under the receiver operating characteristic curve, J number of items, M number of response categories per item

including the Eating Disorder Inventory 2 (EDI-2) [23], at the initiation of nutritional and psychotherapeutic care (t_1) and one year later (t_2). The EDI-2 is an 11-domain scale translated into French and validated by Archinard et al. [24]. We focused on the "Drive for thinness" domain since clinicians felt that the corresponding items could potentially be affected by recalibration after care (recalibration being part of the goals of care). "Drive for thinness" includes 7 items with a six-point Likert response scale ranging from "never" to "always" (1. *I eat sweets and carbohydrates without feeling nervous*; 2. *I think about dieting*; 3. *I feel extremely guilty after overeating*; 4. *I am terrified of gaining weight*;

5. *I exaggerate or magnify the importance of weight*; 6. *I am preoccupied with the desire to be thinner*; 7. *If I gain a pound, I worry that I will keep gaining*). We only considered patients who fully completed the scale at both time points (209 patients of the 210 who attended both visits). We computed the change in the number of GEs (index I) for each patient. Then, we assessed the association between index I and the covariates collected at baseline that seemed relevant to clinicians involved in the cohort (i.e. sociodemographic data, ED characteristics, psychiatric comorbidities [20–22], and character traits measured by the Temperament and Character Inventory [25–27]). We assessed the associations between index I and the other covariates with Mann–Whitney and Kruskal–Wallis tests (for categorical covariates) and Spearman's rank correlation coefficient r (for quantitative covariates). Due to the low discriminating ability of index I, no strong association was expected.

Across the sample, the average age at baseline was 24.2 (s.d. = 8.8), and most patients were women (93%). Of the 209 patients, 31% suffered from restricting anorexia nervosa, 13% from binge eating/purging anorexia nervosa, 25% had bulimia nervosa, 6% displayed binge eating disorder and 24% had eating disorders not otherwise specified (i.e., did not meet the criteria for other diagnoses). The ED had, on average, started 7.3 years before (s.d. = 8.2), and the average BMI was 18.7 kg/m^2 (s.d. = 5.4).

Index I was associated with baseline "Drive for thinness" score ($r = 0.28$, p-value < 0.001) and ideal BMI ($r = -0.17$, p-value $= 0.017$). These results indicated that higher values of index I are associated with greater concern about body image, weight, and shape. In addition, patients with current mood disorders at baseline showed lower distribution of index I than patients without (median = 0 vs 1.5, p-value = 0.036). No other significant association was noticed (Table 2). These results may suggest several clinical hypotheses related to RS. Indeed, patients without mood disorders might be more receptive to interventions targeting cognitive distortions related to body image, and they may therefore be more prone to RS. In addition, among patients with high concern about body image, weight, and shape, care might tend to focus more on deconstructing cognitive distortions and thus induce RS. However, these associations remain globally weak in our sample.

Discussion

Main results

Some evidence of the link between recalibration RS and the change over time in the number of GEs was found in our simulation study. Indeed, as expected, the GEs-based index I remained on average close to 0 among simulated patients not affected by RS, while its average values increased among simulated patients with recalibration. However, the performance of I depended on M, J_{RS}, and the type of recalibration.

The best results were obtained within scenarios with UR. Indeed, when $M > 4$, substantial differences between the means of I among simulated patients with/without recalibration were noticed. These differences were larger when J and J_{RS} were large. When $M = 4$, the index I had lower performance: differences between the means of I among simulated patients with/without RS were small or even nonexistent.

This might be due to the difficulty parameters of the LPCM used to simulate data. Indeed, when $M = 4$, they were widely spaced, and shifts over time (among individuals with recalibration) did not impact the ordering of difficulty parameters. Thus, the difficulty order observed at t_1 still fitted the responses at t_2 of patients with recalibration, resulting in a stable number of GEs over time. However, within scenarios with $M = 7$ or 10, gaps between the difficulty parameters narrowed. Therefore, in these scenarios, shifts over time among individuals with recalibration did impact the ordering of the difficulty parameters (leading to an increase in the number of Guttman errors for these simulated patients). Size of UR used for the simulations might not be detectable with index I when $M = 4$. This issue is problematic and limiting since several domains within QoL questionnaires (SF-36, QLQ-C30…) are composed of items with four or fewer response categories [30, 31]. Additional UR sizes should hence be explored.

When NUR was simulated, differences between the means of I among simulated patients with and without recalibration RS were less marked than with UR. Trends noticed with UR were still observed but less pronounced. Several reasons might explain these results. First, when $M = 4$, the argument evoked for UR concerning the widely spaced difficulty parameters also applies. In addition, we operationalized NUR by differentially shifting the difficulty parameters of the response categories above "1". Thus, the response category "1" of items affected by recalibration kept the same difficulty parameter over time, and some of the shifts for categories above "1" were very small. It might also have hampered the generation of discrepancies at t_2. In addition, unfortunately, the random selection of items affected by recalibration led to shift response categories among the most difficult to make them even more difficult. Again, this has probably hampered the generation of discrepancies at t_2. The effect of the position of the response categories affected by recalibration should hence be explored.

Table 2 Association between index *I* (the change in the number of Guttman errors) and baseline characteristics ($N = 209$)

	Index *I*		
	Categorical variable: median (Q1; Q3) Quantitative variable: Spearman's r	NA	*p*-value
Socio-demographic			
Gender		0	0.770
Female ($n = 195$)	1.2 (−11.0; 17.0)		
Male ($n = 14$)	2.0 (−28.0; 19.5)		
Age (years)	−0.01	0	0.865
Eating disorder characteristics			
Type		0	0.930
AN-R ($n = 65$)	0.0 (−14.0; 20.0)		
AN-BP ($n = 27$)	2.0 (−8.0; 17.0)		
BN ($n = 53$)	5.0 (−11.0; 15.0)		
BED ($n = 13$)	−7.0 (−8.0; 16.0)		
EDNOS ($n = 51$)	2.0 (−12.5; 16.5)		
Duration (years)	−0.06	0	0.368
Lowest BMI (kg/m^2)	−0.05	0	0.439
Ideal BMI (kg/m^2)	−0.17	8	0.017
Current BMI (kg/m^2)	−0.04	1	0.554
Drive for thinness score[a]	0.28		<0.001
Body shape concerns[b]		0	0.070
No to moderate body shape concern ($n = 122$)	−4.0 (−15.5; 16.5)		
Marked body shape concerns ($n = 87$)	5.0 (−5.5; 18.0)		
Psychiatric comorbidities			
Current anxiety disorder		0	0.914
No ($n = 137$)	0.0 (−13.5; 20.5)		
Yes ($n = 72$)	2.0 (−11.0; 15.0)		
Current mood disorder		0	0.036
No ($n = 86$)	0.0 (−12.0; 17.0)		
Yes ($n = 123$)	1.5 (−11.5; 17.5)		
Self-reported character trait[c]			
Cooperativeness	0.01	0	0.869
Self-transcendence	0.03	0	0.671
Self-Directedness	0.03	0	0.710

NA number of missing observations, *Q1* first quartile, *Q3* third quartile, *n* number of patients, *AN-R* anorexia nervosa restricting subtype, *AN-BP* anorexia nervosa binge eating or purging subtype, *BN* bulimia nervosa, *BED* binge eating disorder, *EDNOS* eating disorders not otherwise specified, *BMI* body mass index

[a]A high score indicates a strong search for thinness

[b]Evaluated by the BSQ: Body Shape Questionnaire [28, 29]

[c]Measured by the Temperament and Character Inventory, a high score indicates a more pronounced character trait

Limitations and perspectives

In the simulation study, we decided to focus on recalibration to determine whether index *I* is sensitive to this RS manifestations. However, shifts in GEs could be the result of other types of RS (i.e. reprioritization and reconceptualization), thus further investigations are needed to determine if index *I* is sensitive or insensitive to other RS manifestations.

In addition, we assumed that changes over time in the number of GEs were due to RS, but phenomena other than RS can also interfere in the real world. For instance, a change in the individual latent variable level can impact the range of the possible values for the number of GEs and

hence potentially interfere with index I. Within the simulation study, three configurations were considered regarding the mean change in the latent variable level over time (no change in the average latent variable level, an average decrease of 0.2 in the latent variable level and an average increase of 0.2). The results were very similar among these 3 conditions, but it would be worth investigating larger size of change. The normed number of GEs (i.e., the number of GEs divided by the maximum number of GEs that was achievable given the patient's score and the difficulty order considered) [18] could also be a path to follow to take into account changes in the latent variable at the individual level; the index would hence be the change in the normed number of GEs (denoted I_{norm}). The results for this index within the scenarios emphasized in this article are given in Online Resource 2. It is important to note that we remained under the situation where the questionnaire was still adapted to the population at the second time point. If it turns out that the questionnaire is no longer adapted to the studied population at the second time point, this method would not be adequate (the same would be true for other RS detection methods).

Moreover, phenomena other than RS, such as differential item functioning and violation of the local independence assumption, can also interfere with the index. Indeed, if some patients perceive items differently than the majority of the sample at t_1 (interpreted as differential item functioning, DIF [32, 33]), their responses might result in numerous GEs from the very first measurement occasion. In this case, we may wonder if the changes in the number of GEs is due to RS, DIF, or another phenomenon. In addition, we assumed within the simulation study that the assumption of local independence (at a time point and across time t_1 and time t_2) holds. However, several forms of violation can occur in the real world. For instance, at one time point, two types of violations can occur. First, the targeted latent variable alone may not be sufficient for explaining the correlations among some subsets of items. This violation is referred to as *trait dependence* and is a type of dimensionality issue [34] because additional unmodeled latent variables are involved. Second, the response to one item may depend on the response given to another item. Such an issue is a violation of statistical independence and is referred to as *response dependence* [34]. Both phenomena might impact the number of GEs but in opposite directions. Response dependence, by increasing the similarity of the individuals' responses, might induce a decrease in the number of GEs. In contrast, trait dependence, as an additional source of variation, might induce an increase in the number of GEs [35]. However, these violations of local independence are likely to occur at both time points, limiting the impact on the index. Several diagnosis and detection methods for local dependence exist within Rasch measurement theory (see for instance [35–37]), item response theory [38–44] and the nonparametric item response theory framework [45]. In addition, across measurement occasions, the correlations among an individual's responses might be more important than what the latent variable can explain. For instance, it can be easier to endorse an item when it has already been endorsed before. This phenomenon is also a violation of local independence (response dependence across time points). Olsbjerg and Christensen argued that such a violation could lead to spurious evidence of recalibration and reprioritization (designated by the term "item parameter drift" in their work) [46]. Local dependence across time points has been operationalized as changes in item difficulty parameters over time, depending on the responses given at the first time point [46, 47]. Following this operationalization, violation of local independence could also lead to changes in the number of GEs, resulting in the same manifestation as that for RS. SAS macros, which are available to test the assumption of local independence across time points and item parameter invariance over time within IRT and RMT models at the sample level (based on likelihood ratio tests) [48, 49], can be used to test these assumptions.

Intermittent missing data (MD) were left out of this simulation study as the number of GEs cannot be determined for patients who did not respond to all items. Intermittent MDs are, however, commonly encountered in clinical research and psychometrics. In this case, the normed number of GEs could be computed on the subset of nonmissing items for each patient.

Finally, we simulated samples that were partly affected by only one type of recalibration at a time (UR or NUR) with the same size of RS for all affected patients. The individual nature of this phenomenon was neglected. Simulations with subgroups of patients affected by different sizes and types of recalibration should be explored.

Alone, the change in the number of GEs has a low discriminating ability. However, it could be used as a preliminary analysis (when RS occurrence is suspected) to identify covariates associated with RS or possibly a subgroup of patients more likely to present RS. Therefore, index I may guide the choice of the adequate method for identifying RS and estimate its size (by introducing a covariate or conducting the analysis at the subgroup level). However, the methodology for defining the threshold to classify individuals must still be developed. Indeed, we have shown in our simulation study that patients without RS had a value of I that fluctuated on average around approximately 0, yet some variability was observed, notably when J and M increased. This variability generated an overlap in the distributions of index I for patients with and without RS. This phenomenon makes it difficult to define a threshold for the index (which would likely be a function of J and

M). To overcome the effect of the questionnaire structure on the threshold, the normed number of GEs could be used instead of the number of GEs since it takes into account the maximum number of GEs reachable for each individual given the questionnaire structure.

Conclusion

Some evidence of the link between RS and the change in GEs was found in this study. GEs could be a valuable non-parametric tool for RS detection at a more individual level, but further investigation is needed.

Appendix 1: simulation implementation

Longitudinal partial credit model

The longitudinal Partial Credit Model (LPCM) was chosen to generate data since it allowed modelling response categories probabilities of polytomous items forming a unidimensional scale across time, and provided a possibility to simulate RS for a changing proportion of patients. The probability of patient n to answer m ($= 0, \ldots, M-1$) on item j at time t under the LPCM is given by:

$$P\left(X_{nj}^{(t)} = m|\theta_n^{(t)}, \delta_{j1}^{(t)}, \ldots, \delta_{jM-1}^{(t)}\right) = \frac{\exp\left(m.\theta_n^{(t)} - \sum_{p=1}^{m} \delta_{jp}^{(t)}\right)}{\sum_{l=0}^{M-1} \exp\left(l.\theta_n^{(t)} - \sum_{p=1}^{l} \delta_{jp}^{(t)}\right)}$$

where $X_{nj}^{(t)}$ denotes the response to the item $j = 1, \ldots, J$ of the individual n at time t

$\theta_n^{(t)}$ stands for the latent variable level of the individual n at t (realization of the random variable Θ).

$$\begin{pmatrix} \Theta^{(t_1)} \\ \Theta^{(t_2)} \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{bmatrix}\right)$$

$\delta_{jm}^{(t)}$ is the difficulty of the response category $m = 1, \ldots, M-1$ from item j at the time point t. If $\delta_{jm}^{(t)}$ is low, the proportion of patients scoring m or more to item j will be high: m is hence an easy response category (vice versa for difficult response categories). Null response categories do not have a difficulty parameter.

At the first measurement occasion, difficulty parameters were chosen to be spaced along the latent variable continuum (assumed normally distributed, with a zero mean and a standard deviation equaled to 1). For each item$j$, the difficulty parameter of the first positive response category (denoted $\delta_{j1}^{(t_1)}$) equaled the $\frac{j}{J+1}th$ quantile from a $N(0,1)$. Difficulty parameters of the following response categories were then regularly shifted from the first one: $\delta_{jm}^{(t_1)} = \delta_{j1}^{(t_1)} + (m-1) \times \frac{2}{M-2}$. Finally, difficulty parameters of all items were centered on the mean $\overline{\delta} = \frac{\sum_{j,m} \delta_{jm}^{(t_1)}}{J(M-1)}$ so that difficulty parameters were centered on the mean of the latent variable distribution (i.e. 0). It hence corresponded to the situation where the questionnaire is suitable for a population with a latent variable following a standard normal distribution. At the first measurement occasion, the model is a rating scale model.

Recalibration operationalization

To simulate the responses of patients affected by UR at t_2, we choose to shift by -1 all the difficulty parameters of the item(s) affected by recalibration, making all response categories easier. For patients affected by NUR, difficulty parameters were differentially shifted by values ranging 0 to $2\,2\eta$, with $\eta = 1.8$: the first positive response category kept the same difficulty parameter over time, while other categories became more difficult. Finally, we kept the difficulty parameters constant over time to simulate the responses of patients not affected by RS.

For all m in $\{1, \ldots, M-1\}$, $\delta_{jm}^{(t_2)} = \begin{cases} \delta_{jm}^{(t_1)} + \eta_m & \text{for individuals affected by RS} \\ \delta_{jm}^{(t_1)} & \text{for individuals not affected by RS} \end{cases}$

For UR, $\eta_m^{UR} = -1$ for all m in $\{1, \ldots, M-1\}$

For NUR,

$$\eta_m^{NUR} = \begin{cases} \frac{(m-1)\eta}{m} & \text{if } 1 \le m < \frac{M}{2} \\ \eta & \text{if } m = \frac{M}{2} \\ \frac{(M-m+1)\eta}{M-m} & \text{if } \frac{M}{2} < m \le M-1 \end{cases} \quad \text{where } \eta = 1.8$$

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11136-021-03015-9.

Acknowledgements We would like to warmly thank all the staff members of the EVALADD cohort.

Funding Y. Dubuy received a national grant from the French Ministry of Higher Education, Research and Innovation. The EVALADD cohort is sponsored by Nantes University Hospital (CHU Nantes).

Data availability Modules, scripts and an extract of the simulated data used in the paper are available at the Open Science Framework via the link: https://osf.io/h9nyd/?view_only=b196db78f31c4e9fbb07013342a133a2

Declarations

Conflict of interest Authors declare that they have no conflict of interest.

Ethical approval The EVALADD cohort (Investigator: M. Grall-Bronnec) was approved by the local Research Ethics Committee (Groupe Nantais d'Ethique dans le Domaine de la Santé), by the CCTIRS (Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le domaine de la Santé) and by the CNIL (Commission Nationale de l'Informatique et des Libertés).

Informed consent All participants provided written informed consent (for under 18-year-olds, a legal representative provided informed consent), in accordance with the Helsinki declaration.

References

1. Basch, E. (2017). Patient-reported outcomes: Harnessing patients' voices to improve clinical care. *The New England Journal of Medicine, 376*(2), 105–108. https://doi.org/10.1056/NEJMp1611252
2. Schwartz, C. E., Finkelstein, J. A., & Rapkin, B. D. (2017). Appraisal assessment in patient-reported outcome research: Methods for uncovering the personal context and meaning of quality of life. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 26*(3), 545–554. https://doi.org/10.1007/s11136-016-1476-2
3. Sprangers, M. A. G., & Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: A theoretical model. *Social Science & Medicine, 48*(11), 1507–1515. https://doi.org/10.1016/S0277-9536(99)00045-3
4. Vanier, A., Falissard, B., Sébille, V., & Hardouin, J. B. (2017). The complexity of interpreting changes observed over time in health-related quality of life: A short overview of 15 years of research on response shift theory. In F. Guillemin, A. Leplege, S. Briancon, E. Spitz, & J. Coste (Eds.), *Perceived health and adaptation in chronic disease* (1st ed.). New York: Routledge.
5. Schwartz, C. E., Sprangers, M. A., & Fayers, P. M. (2005). *Response shift: You know it's there, but how do you capture it? Challenges for the next phase of research. In Assessing quality of life in clinical trials* (2nd ed.). Oxford University Press.
6. Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 14*(3), 587–598.
7. Schwartz, C. E. (2016). Introduction to special section on response shift at the item level. *Quality of Life Research, 25*(6), 1323–1325. https://doi.org/10.1007/s11136-016-1299-1
8. Guilleux, A., Blanchin, M., Vanier, A., Guillemin, F., Falissard, B., Schwartz, C. E., Hardouin, J. B., & Sébille, V. (2015). RespOnse Shift ALgorithm in Item response theory (ROSALI) for response shift detection with missing data in longitudinal patient-reported outcome studies. *Quality of Life Research, 24*(3), 553–564. https://doi.org/10.1007/s11136-014-0876-4
9. Blanchin, M., Guilleux, A., Hardouin, J.-B., & Sébille, V. (2020). Comparison of structural equation modelling, item response theory and Rasch measurement theory-based methods for response shift detection at item level: A simulation study. *Statistical Methods in Medical Research, 29*(4), 1015–1029. https://doi.org/10.1177/0962280219884574
10. Vanier, A., Sébille, V., Blanchin, M., Guilleux, A., & Hardouin, J.-B. (2015). Overall performance of Oort's procedure for response shift detection at item level: A pilot simulation study. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 24*(8), 1799–1807. https://doi.org/10.1007/s11136-015-0938-2
11. Nolte, S., Mierke, A., Fischer, H. F., & Rose, M. (2016). On the validity of measuring change over time in routine clinical assessment: A close examination of item-level response shifts in psychosomatic inpatients. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 25*(6), 1339–1347. https://doi.org/10.1007/s11136-015-1123-3
12. Gandhi, P. K., Schwartz, C. E., Reeve, B. B., DeWalt, D. A., Gross, H. E., & Huang, I.-C. (2016). An item-level response shift study on the change of health state with the rating of asthma-specific quality of life: A report from the PROMIS(®) Pediatric Asthma Study. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 25*(6), 1349–1359. https://doi.org/10.1007/s11136-016-1290-x
13. Verdam, M. G. E., Oort, F. J., & Sprangers, M. A. G. (2016). Using structural equation modeling to detect response shifts and true change in discrete variables: An application to the items of the SF-36. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 25*(6), 1361–1383. https://doi.org/10.1007/s11136-015-1195-0
14. Ahmed, S., Sawatzky, R., Levesque, J.-F., Ehrmann-Feldman, D., & Schwartz, C. E. (2014). Minimal evidence of response shift in the absence of a catalyst. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 23*(9), 2421–2430. https://doi.org/10.1007/s11136-014-0699-3
15. Blanchin, M., Sébille, V., Guilleux, A., & Hardouin, J.-B. (2016). The Guttman errors as a tool for response shift detection at subgroup and item levels. *Quality of Life Research, 25*(6), 1385–1393. https://doi.org/10.1007/s11136-016-1268-8
16. Meijer, R. R., Niessen, A. S. M., & Tendeiro, J. N. (2016). A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics: Examples and a

computer program. *Assessment, 23*(1), 52–62. https://doi.org/10.1177/1073191115577800
17. Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. SAGE.
18. Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement, 32*(3), 224–247. https://doi.org/10.1177/0146621607302479
19. Fischer, G. H., & Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika, 59*(2), 177–192. https://doi.org/10.1007/BF02295182
20. American Psychiatric Association, & American Psychiatric Association (eds.). (2009). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR* (4. ed., text revision, 13. print.). Arlington, VA: American Psychiatric Assoc.
21. Lecrubier, Y., Sheehan, D., Weiller, E., Amorim, P., Bonora, I., Harnett Sheehan, K., & Dunbar, G. (1997). The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: Reliability and validity according to the CIDI. *European Psychiatry, 12*(5), 224–231. https://doi.org/10.1016/S0924-9338(97)83296-8
22. Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., & Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of Clinical Psychiatry, 59*(Suppl 20), 22–33.
23. Garner, D. M. (1991). *Eating disorder inventory-2. Professional manual*. Psychological Assessment Research.
24. Archinard, M., Rouget, P., Painot, D., & Liengme, C. (2002). Inventaire des troubles alimentaires 2 [Eating Disorder Inventory 2]. In M. Bouvard & J. Cottraux (Eds.), *Protocoles et échelles d'évaluation en psychiatrie et en psychologie [Protocols and evaluation scales in psychiatry and psychology]* (3rd ed., pp. 249–251). Masson.
25. Cloninger, C. R., Przybeck, T. R., & Svrakic, D. M. (1994). *The temperament and character inventory (TCI) a guide to its development and use*. Center for Psychobiology of Personality, Washington University.
26. Pélissolo, A., & Lépine, J.-P. (1997). Traduction française et premières études de validation du questionnaire de personnalité TCI. [Validation study of the French version of the TCI.]. *Annales Médico-Psychologiques, 155*(8), 497–508.
27. Chakroun-Vinciguerra, N., Faytout, M., Pélissolo, A., & Swendsen, J. (2005). Validation française de la version courte de l'Inventaire du Tempérament et du Caractère (TCI-125). *Journal de Thérapie Comportementale et Cognitive, 15*(1), 27–33. https://doi.org/10.1016/S1155-1704(05)81209-1
28. Cooper, P. J., Taylor, M. J., Cooper, Z., & Fairbum, C. G. (1987). The development and validation of the body shape questionnaire. *International Journal of Eating Disorders, 6*(4), 485–494. https://doi.org/10.1002/1098-108X(198707)6:4%3c485::AID-EAT2260060405%3e3.0.CO;2-O
29. Rousseau, A., Knotter, A., Barbe, P., Raich, R., & Chabrol, H. (2005). Validation of the French version of the Body Shape Questionnaire. *L'Encephale, 31*(2), 162–173. https://doi.org/10.1016/s0013-7006(05)82383-8
30. Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care, 30*(6), 473–483.
31. Aaronson, N. K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N. J., & de Haes, J. C. (1993). The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute, 85*(5), 365–376. https://doi.org/10.1093/jnci/85.5.365
32. Holland, P. W., & Wainer, H. (Eds.). (1993). Differential item functioning. *Differential Item Functioning*, xv, 453–xv, 453.
33. Osterlind, S., & Everson, H. (2009). *Differential item functioning*. SAGE Publications.
34. Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*(3), 200–215.
35. Christensen, K. B., Kreiner, S., & Mesbah, M. (Eds.). (2013). *Rasch models in health*. ISTE.
36. Andrich, D., & Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the rasch model. *Applied Psychological Measurement, 34*(3), 181–192. https://doi.org/10.1177/0146621609360202
37. Andrich, D., Humphry, S. M., & Marais, I. (2012). Quantifying local, response dependence between two polytomous items using the Rasch model. *Applied Psychological Measurement, 36*(4), 309–324. https://doi.org/10.1177/0146621612441858
38. Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125–145. https://doi.org/10.1177/014662168400800201
39. Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265. https://doi.org/10.2307/1165285
40. Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods, 2*(3), 261–277. https://doi.org/10.1037/1082-989X.2.3.261
41. Douglas, J., Kim, H. R., Habing, B., & Gao, F. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics, 23*(2), 129–151. https://doi.org/10.2307/1165318
42. Ip, E. H. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika, 66*(1), 109–132. https://doi.org/10.1007/BF02295736
43. Ip, E. H. (2002). Locally dependent latent trait model and the dutch identity revisited. *Psychometrika, 67*(3), 367–386. https://doi.org/10.1007/BF02294990
44. Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods, 23*(1), 138–149. https://doi.org/10.1037/met0000121
45. Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2016). Using conditional association to identify locally independent item sets. *Methodology, 12*(4), 117–123. https://doi.org/10.1027/1614-2241/a000115
46. Olsbjerg, M., & Christensen, K. B. (2015). Modeling local dependence in longitudinal IRT models. *Behavior Research Methods, 47*(4), 1413–1424. https://doi.org/10.3758/s13428-014-0553-0
47. Marais, I. (2009). Response dependence and the measurement of change. *Journal of Applied Measurement, 10*, 17–29.
48. Olsbjerg, M., & Christensen, K. B. (n.d.). LIRT: SAS macros for longitudinal IRT models, 49.
49. Olsbjerg, M., & Christensen, K. B. (2015). %lrasch_mml : A *SAS* macro for marginal maximum likelihood estimation in longitudinal polytomous rasch models. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v067.c02

Evaluation of the link between the Guttman errors and response shift at the individual level:

Electronic supplementary materials #1

Yseulys Dubuy, Véronique Sébille, Marie Grall-Bronnec, Gaëlle Challet-Bouju, Myriam Blanchin, Jean-Benoit Hardouin

Caption:

This online resource provides the results obtained with the change in the number of Guttman errors over time within all the scenarios considered (means among patients with/without RS and corresponding AUROCs). The change in the number of Guttman errors over time is denoted *I*. Results are given for uniform recalibration and then for non-uniform recalibration.

Note de rédaction de thèse :

Ce fichier supplémentaire n'a pas été inclus dans le manuscrit de thèse, car il s'agit d'un listing de résultats très similaires à ceux figurant déjà dans l'article. Il est néanmoins accessible en ligne à partir du lien donné dans la section ***"Supplementary information"*** *de l'article.*

Evaluation of the link between the Guttman errors and response shift at the individual level:

Electronic supplementary materials #2

Yseulys Dubuy, Véronique Sébille, Marie Grall-Bronnec, Gaëlle Challet-Bouju, Myriam Blanchin, Jean-Benoit Hardouin

Caption:

This online resource provides the results obtained with the change in the normed number of Guttman errors over time within the subset of scenarios emphasized in the article: N = 200 (sample size), $p = 25\%$ (proportion of patients affected by response shift), $\Delta = -0.2$ (average change in the latent variable over time), uniform recalibration or non-uniform recalibration. The change in the normed number of Guttman errors over time is denoted I_{norm}.

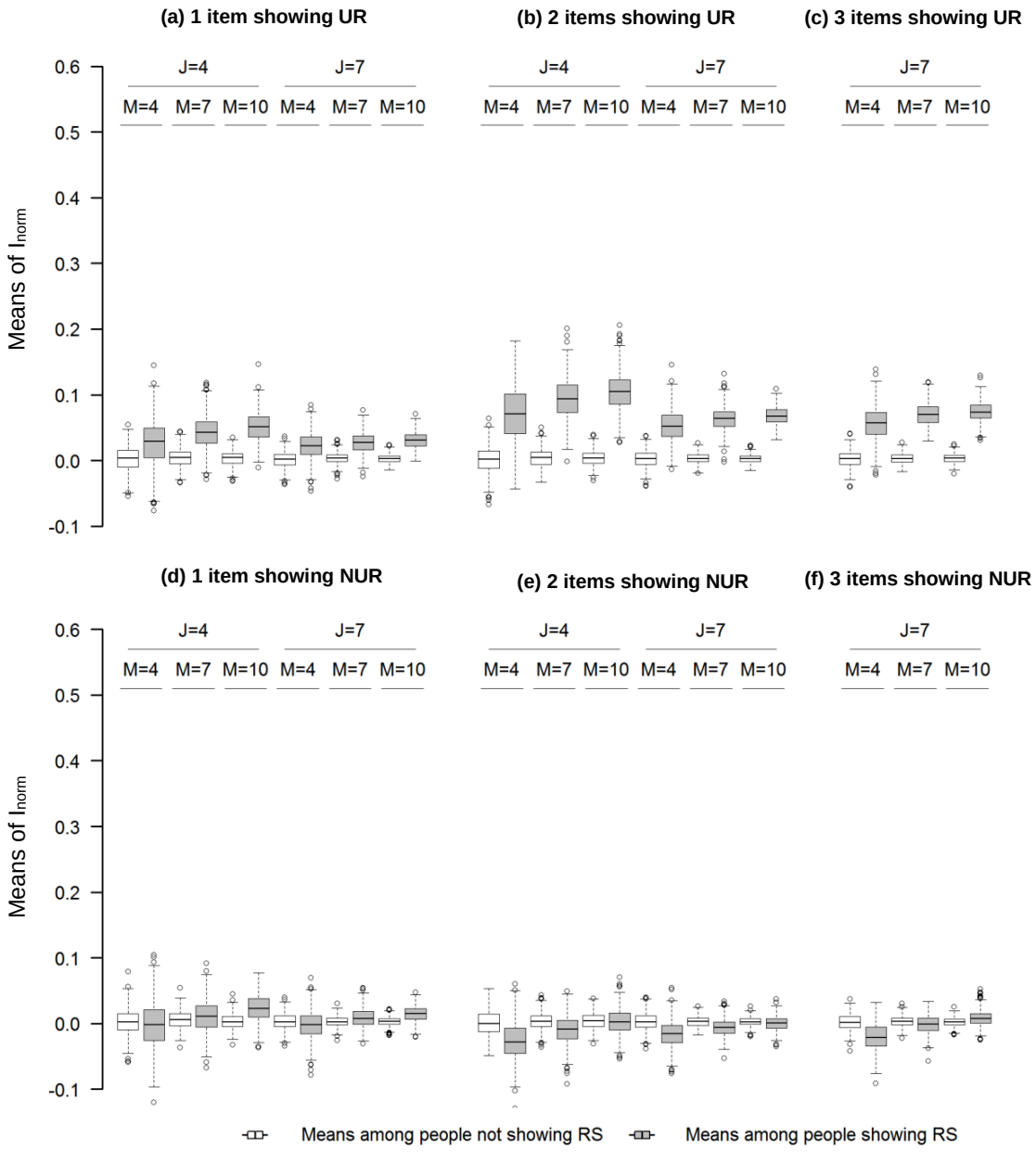


Fig 1 Boxplots of the 500 mean values of index I_{norm} obtained for each scenario among people affected by response shift (in white) and among people not affected by response shift (in grey). Each pair of boxplots corresponds to one scenario. UR: uniform recalibration; NUR: non-uniform recalibration; J: number of items; M: number of response categories per item**.** Subset of scenarios considered: N = 200 (sample size), $p = 25\%$ (proportion of patients affected by response shift), $\Delta = -0.2$ (average change in the latent variable over time)

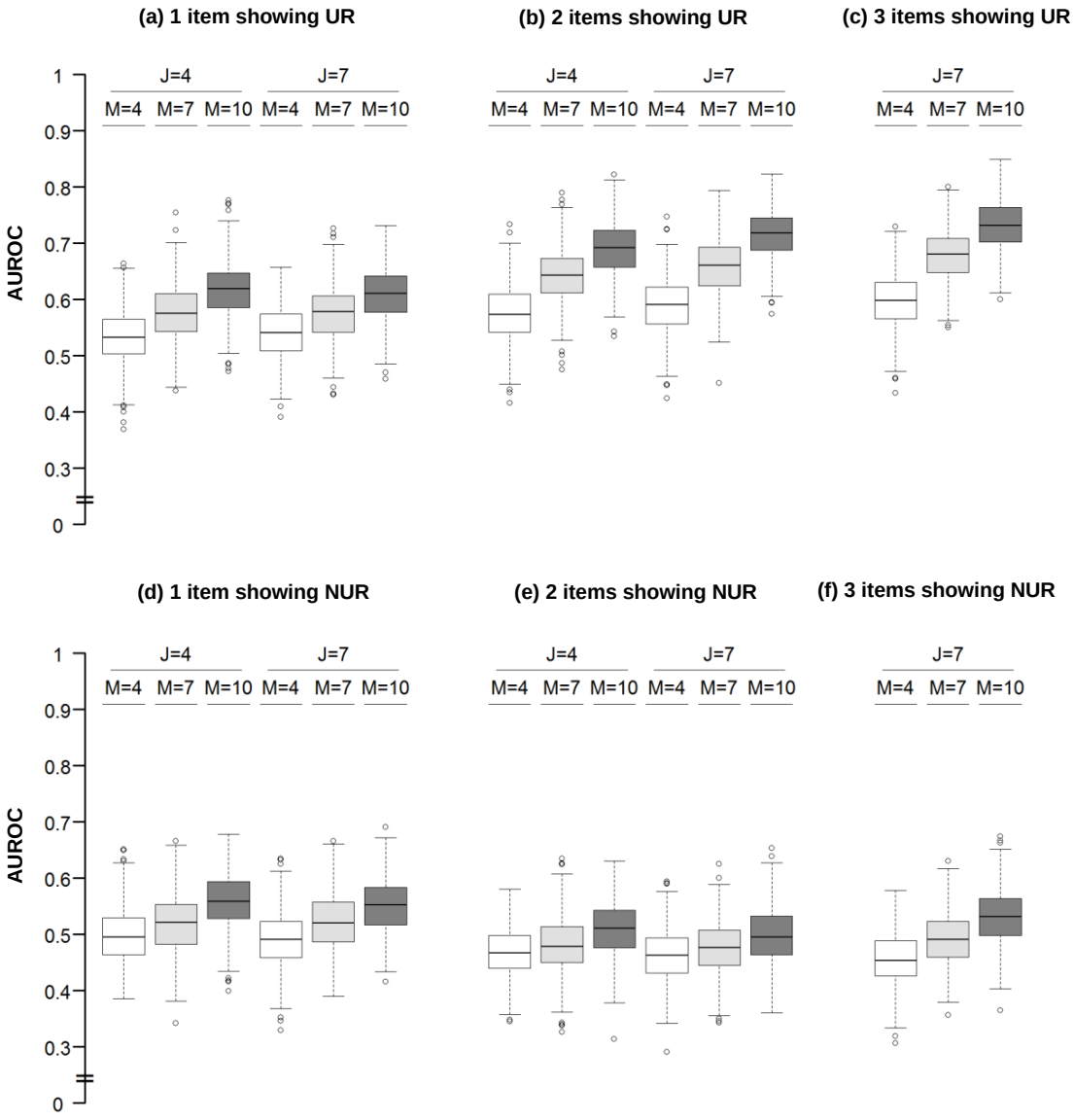


Fig 2 Boxplots of the 500 AUROCs associated with index I_{norm} obtained for each scenario. UR: uniform recalibration; NUR: non-uniform recalibration; J: number of items; M: number of response categories per item**.** Subset of scenarios considered: N = 200 (sample size), $p = 25\%$ (proportion of patients affected by response shift), $\Delta = -0.2$ (average change in the latent variable over time)

Evaluation of the link between the Guttman errors and response shift at the individual level:

Additional file: Description of the Stata module `eguttman` created for the simulation study

Yseulys Dubuy, Véronique Sébille, Marie Grall-Bronnec, Gaëlle Challet-Bouju, Myriam Blanchin, Jean-Benoit Hardouin

Description

The `eguttman` module computes the (normed) number of Guttman errors for each individual of a cross-sectional or longitudinal dataset. This module can handle dichotomous and polytomous items. The user must define: (i) the items among which Guttman errors will be counted, (ii) the name of the column (i.e., the name of the variable) containing the measurement occasions, (iii) the measurement occasion considered to determine the difficulty order of the response categories, (iv) the name for the variables created by the module, and (v) the maximal number of response categories for the items.

Syntax

`eguttman` *varlist*`, visit(`*varname*`) reference(`*string*`) generror(`*string*`) respmax(`*int*`) [genscore(`*string*`) replace]`

varlist contains the variables (i.e., the items) used to count the Guttman errors.

`visit(`*varname*`)` specifies the name of the variable containing the measurement occasions. This variable must be categorical. For instance, it can contain: V1, V2, V3, etc. or T1, T2, T3, etc.
If the data set is cross-sectional, a variable containing the same category for each observation of the data set must be created.

`reference(`*string*`)` allows the user to specify the measurement occasion based on which the difficulty order will be defined. The string entered must correspond to one of the categories of the variable specified in `visit`.

`generror(`*string*`)` enables the user to specify the names of the variables created by the `eguttman` module. Specifically, if the user chooses to enter generror(NB_GE), then the module will create the following variables:

- NB_GE: the number of Guttman errors for each individual and each measurement occasion
- MAX_NB_GE: the maximal number of Guttman errors reachable for each individual given his/her score, computed based on Emons' algorithm [1] (implemented directly into the module)
- NORM_NB_GE: the normed number of Guttman errors for each individual (NB_GE/MAX_NB_GE), computed based on Emons' algorithm [1]

Notes:

- The Guttman errors are defined based on the difficulty order determined at the measurement occasion chosen as reference.
- If the score is null or maximal, there will be no Guttman errors possible. Hence the normed number of Guttman errors is set to zero. Besides,

`respmax(`*int*`)` contains the maximal response categories of the items. Of note, this version of the module assumes that: (i) the first response category of all items is coded "0", and (ii) all items have the same number of response categories.

`genscore(`*string*`):` if the option is fulfilled, then the module `eguttman` will create a variable containing the sum of the item responses indicated in *varlist*. This variable will be named as the *string* indicated in the option.

`replace:` if the option is indicated, then the module will overwrite the variables already existing in the data set by new ones.

Important considerations

- The version disseminated on OSF has not been designed to deal with missing data.
- If there are ties in the difficulty order, the module will consider that no Guttman errors are possible between the response categories involved in each group of ties.
- If the data are longitudinal (several measurement occasions), then the dataset must be in long-format .

Example with cross-sectional data

This example requires the Stata module `simirt` [2] to simulate the data.

```
/* Step 1: Simulation of a dataset containing the responses of 200 individuals responding to a unidimensional questionnaire composed of 4 items (with 4 response categories each, coded 0, 1, 2 and 3). Data are simulated based on a partial credit model and individual latent variable levels are drown from a normal distribution with a mean of 0 and a standard deviation set to 1*/

/* Step 1.1: Creation of a matrix D containing the items thresholds used for the partial credit model */
. mat D= (-1.84,-0.84,0.16 \ -1.25,-0.25,0.75\-0.75,0.25,1.25 \-0.16,0.84,1.84)

/* Step 1.2: Data set generation (items are named item1, item2, item3 and item4) */
. qui simirt, nbobs(200) mu(0) cov(1) dim(4) pcm(D) clear

/* Step 2: Computation of the number of Guttman errors*/

/* Step 2.1: Creation of the variable containing the only measurement occasion, denoted T1*/
. gen time = "T1"

/* Step 2.2: Launch the eguttman module */
. eguttman item*, visit(time) reference(T1) generror(NB_GE) respmax(3)
4 items are taken into account
Visits considered: T1
------------------------------------------------------------------------
Reference modality of time to compute the items difficulties: T1
------------------------------------------------------------------------
Difficulty item1sup2: .27
Difficulty item1sup3: .6
Difficulty item2sup1: .205
Difficulty item2sup2: .405
Difficulty item2sup3: .72
Difficulty item3sup1: .275
Difficulty item3sup2: .625
Difficulty item3sup3: .87
Difficulty item4sup1: .345
Difficulty item4sup2: .71
Difficulty item4sup3: .94
------------------------------------------------------------------------
```

The output summarizes the main information: the number of items in *varlist*, the measurement occasion(s) considered, and the difficulty order observed at time T1. Three variables are created in the dataset for the number of Guttman errors (see the screenshot below).

	id	lt1	item1	item2	item3	item4	time	NB_GE	max_NB_GE	norm_NB_GE
1	1	-1.413241	1	1	0	0	T1	0	11	0
2	2	-.229486	0	1	0	0	T1	1	4	.25
3	3	.726616	3	3	0	1	T1	6	23	.2608696
4	4	1.291213	3	0	1	2	T1	8	28	.2857143

References

[1] Emons WHM. Nonparametric Person-Fit Analysis of Polytomous Item Scores. *Applied Psychological Measurement*. 2008;32(3):224-247. doi:10.1177/0146621607302479

[2] Hardouin JB. "SIMIRT: Stata module to process data generated by IRT models," Statistical Software Components S450402, Boston College Department of Economics

<u>Notes de rédaction de thèse :</u>

Le module a été retravaillé après les études de simulation que nous avons réalisées pour pouvoir :

1. *Utiliser la méthode proposée par Koopman et al. (doi : 10.1007/978-3-319-56294-0_17) pour la gestion des ex-aequo (en plus de la méthode que nous avons utilisé par défaut qui suppose qu'il ne peut pas y avoir d'erreurs de Guttman entre des modalités de réponse ayant la même difficulté).*

2. *Imputer les données manquantes de façon à minimiser le nombre d'erreurs de Guttman « créées » suite à l'imputation.*

4.3.2 Commentaires complémentaires

Pour analyser cette étude de simulation, nous avons initialement choisi d'estimer les moyennes des deux indicateurs I et I_{norm} dans les deux groupes d'individus suivants (pour chacune des réplications) :

- Les individus pour qui de la recalibration avait été simulée (même recalibration pour tous) ;
- Les individus pour qui on n'avait pas simulé de recalibration.

L'objectif était de démontrer que ces indicateurs étaient en moyenne plus élevés chez les individus pour qui on avait simulé de la recalibration. Nous nous sommes également intéressés à la capacité discriminante de ces indicateurs, afin de déterminer s'ils permettaient de distinguer les deux groupes d'individus. Cette capacité discriminante a été estimée grâce à l'aire sous la courbe ROC (*receiver operating characteristic*) non paramétrique associée à ces deux indicateurs. Les moyennes et les aires sous la courbe ROC (AUROC) obtenues pour chaque réplication d'un même scénario ont ensuite été résumées à l'aide de *box plots* (aussi appelés boîtes à moustaches). Ces analyses sont représentées schématiquement par la figure 4.4.

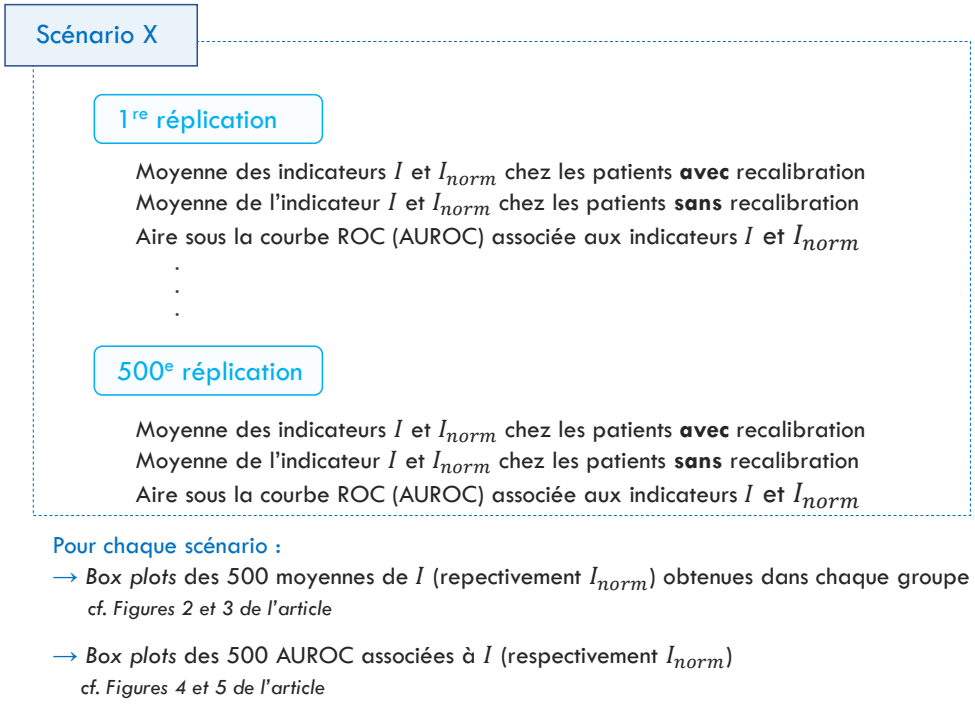


FIGURE 4.4 – Analyse de l'étude de simulation

Pour rappel, la courbe ROC permet de représenter graphiquement la relation entre le taux de vrais positifs (la sensibilité) et le taux de faux positifs (1 - spécificité) des indicateurs étudiés. Pour construire la courbe ROC non paramétrique associée à chaque indicateur sur une réplication précise, les valeurs observées de l'indicateur (notées s) ont été classées par ordre croissant, puis pour chacune de ces valeurs, la sensibilité et la spécificité associées ont été calculées à partir des formules suivantes :

$$sensibilité = \frac{VP}{VP + FN} \qquad spécificité = \frac{VN}{FP + VN}$$

Où VP, FP, VN et FN représentent les effectifs de la matrice de confusion définie pour chaque seuil s :

		Indicateur étudié		
		Indicateur $\leq s$ *Individus peu susceptibles d'expérimenter de la RC*	Indicateur $> s$ *Individus susceptibles d'expérimenter de la RC*	Total
Individus avec RC simulée ?	Non	Vrais négatifs VN	Faux positifs FP	$VN + FP$
	Oui	Faux négatifs FN	Vrais positifs VP	$FN + VP$
	Total	$VN + FN$	$FP + VP$	N

Notes : *RC = Recalibration, N = Effectif total*

Les couples (1 – spécificité, sensibilité) sont ensuite représentés graphiquement, formant ainsi la courbe ROC non paramétrique associée à l'indicateur. Un indicateur discriminant est caractérisé par une courbe ROC s'approchant du point de coordonnées (1,1) et une aire sous courbe ROC proche de 1. Au contraire, un indicateur peu ou pas discriminant est caractérisé par une courbe ROC proche de la diagonale et une aire sous courbe proche de 0,5. (voir figure 4.5 pour une illustration).

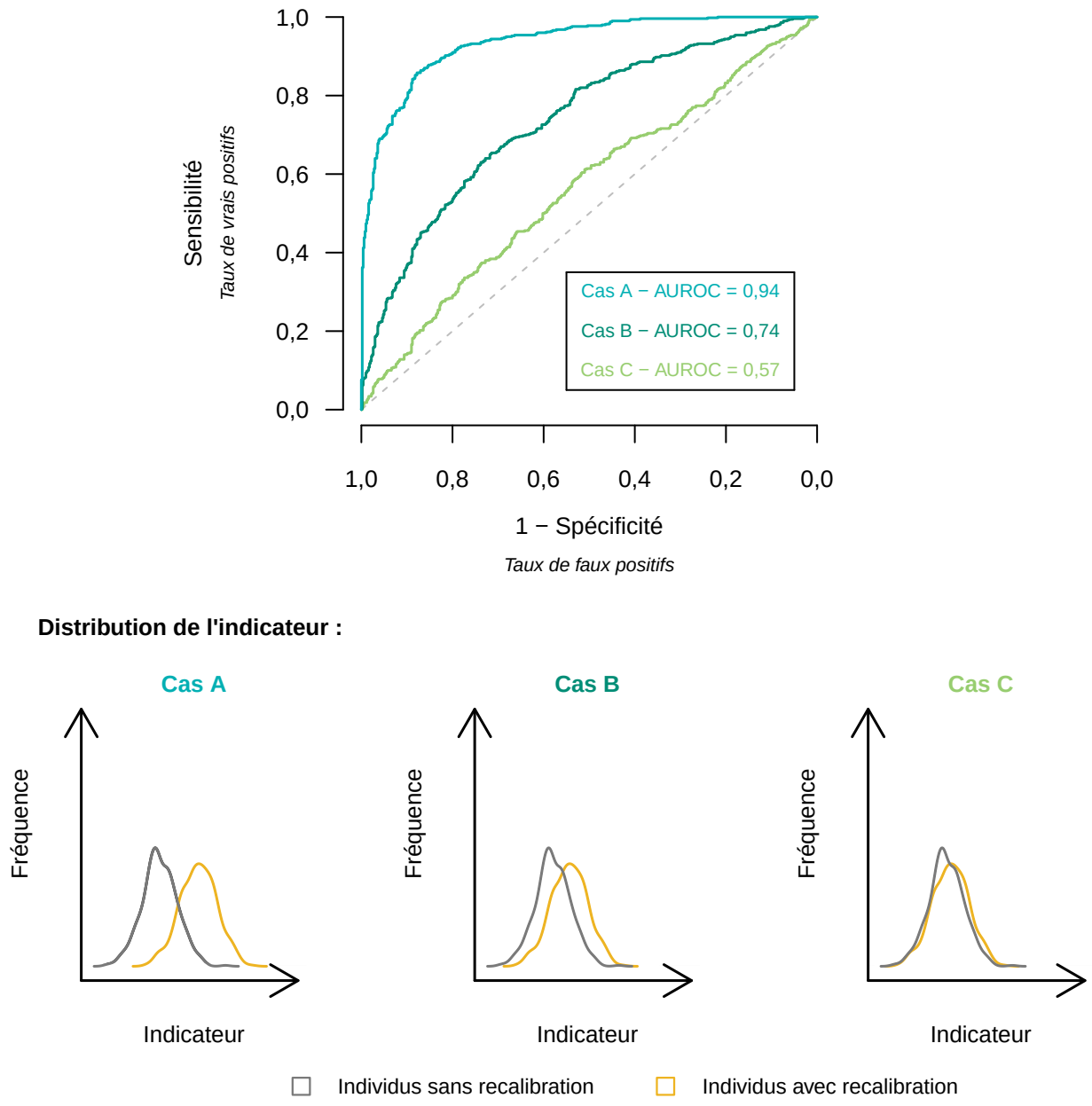


FIGURE 4.5 – Courbes ROC et aires sous courbes (AUROC) pour un indicateur ayant une capacité discriminante de la recalibration variable

Notes :
Cas A – L'indicateur est discriminant : faible chevauchement des distributions de l'indicateur dans les deux groupes (se manifestant par une AUROC proche de 1)
Cas B – Cas intermédiaire
Cas C – L'indicateur n'est pas discriminant : important chevauchement des distributions de l'indicateur dans les deux groupes (se manifestant par une AUROC proche de 0,5)

Le choix de résumer la distribution des deux indicateurs à leurs valeurs moyennes dans les deux groupes d'individus considérés (avec ou sans recalibration simulée) s'est avéré discutable. En effet, cela nous a fait perdre la granularité individuelle que l'on souhaitait investiguer. Il est néanmoins difficile de représenter des données individuelles dans une étude de simulation où tous les scénarios ont été répliqués 500 fois.

De plus, les écarts observés entre les *box plots* des moyennes dans les deux groupes d'individus se sont avérés trompeurs. En effet, dans certains scénarios, les moyennes des deux indicateurs semblaient clairement plus élevées chez les individus avec recalibration que chez les individus sans. La capacité discriminante des indicateurs parmi ces scénarios s'est pourtant avérée faible, avec des AUROC ne dépassant pas 0,7. Ces valeurs témoignent d'un fort chevauchement des distributions des indicateurs dans les groupes à discriminer. Ce chevauchement n'est pas discernable à partir des figures de l'article sur les moyennes des indicateurs (cf. figures 2 et 3 de l'article, pages 123 et 124 et figure 1 du document "*Electronic supplementary materials* #2", page 135). De fait, en ne calculant que la moyenne des distributions, nous avons perdu l'information concernant la dispersion des indicateurs.

Afin d'illustrer cette problématique, les *box plots* des moyennes de l'indicateur I ont été mis en regard des distributions de l'indicateur I dans les deux groupes à discriminer (figure 4.6) pour les scénarios vérifiant les conditions suivantes :

- Taille de l'échantillon : $N = 200$;
- Proportion d'individus pour qui on a simulé de la recalibration : $p = 25\ \%$;
- Nombre d'items dans le questionnaire : $J = 7$ items ;
- Nombre de modalités de réponse pour chaque item : $M = 4$, 7 ou 10 ;
- Variable latente qui se détériore en moyenne entre t_1 et t_2 : $\Delta = -0{,}2$;
- Recalibration uniforme se manifestant (ou touchant) $J_{RS} = 2$ items : les items 6 et 7.

Ces scénarios ont été sélectionnés, car ils permettaient d'illustrer globalement la palette des performances de l'indicateur et qu'ils font partie des scénarios dont les résultats sont présentés dans l'article. Mis à l'échelle de la distribution de l'indicateur I, les écarts entre les *box plots* des moyennes sont moindres. La même figure a également été réalisée pour l'indicateur I_{norm} (figure 4.7). Au vu de ce constat, seules les aires sous la courbe ROC seront considérées pour la suite de ces travaux.

Notes : *Dans les figures 4.6 et 4.7, il n'était pas possible de représenter les distributions des indicateurs obtenues pour chacun des 500 jeux de données associés à un scénario. C'est pourquoi, les distributions des indicateurs sont données : (i) sur l'ensemble des 500 réplications mises en commun (trait plein et épais) et (ii) sur quatre jeux de données choisis au hasard parmi l'ensemble des 500 réplications (traits fins discontinus).*

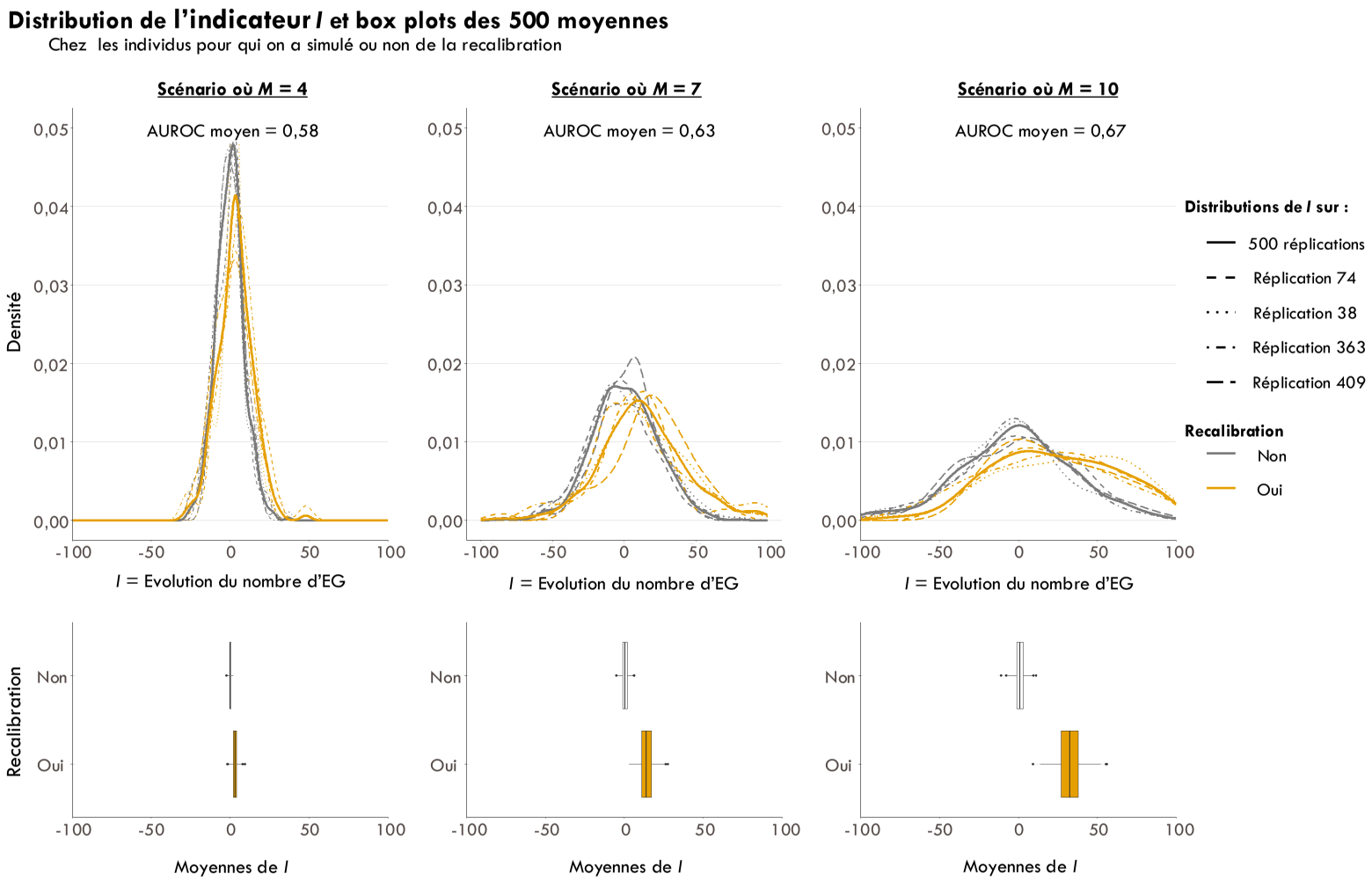


FIGURE 4.6 – *Box plots* des moyennes de **l'indicateur *I*** chez les patients avec et sans recalibration et distribution de l'indicateur dans ces mêmes sous-groupes. Trois scénarios considérés : $N = 200$ individus, $J = 7$ items, $M = 4$, 7 ou 10 modalités de réponse, Recalibration uniforme simulée pour 25% de l'échantillon et touchant $J_{RS} = 2$ items, Moyenne de la variable latente qui se détériore en moyenne ($\Delta = -0,2$)

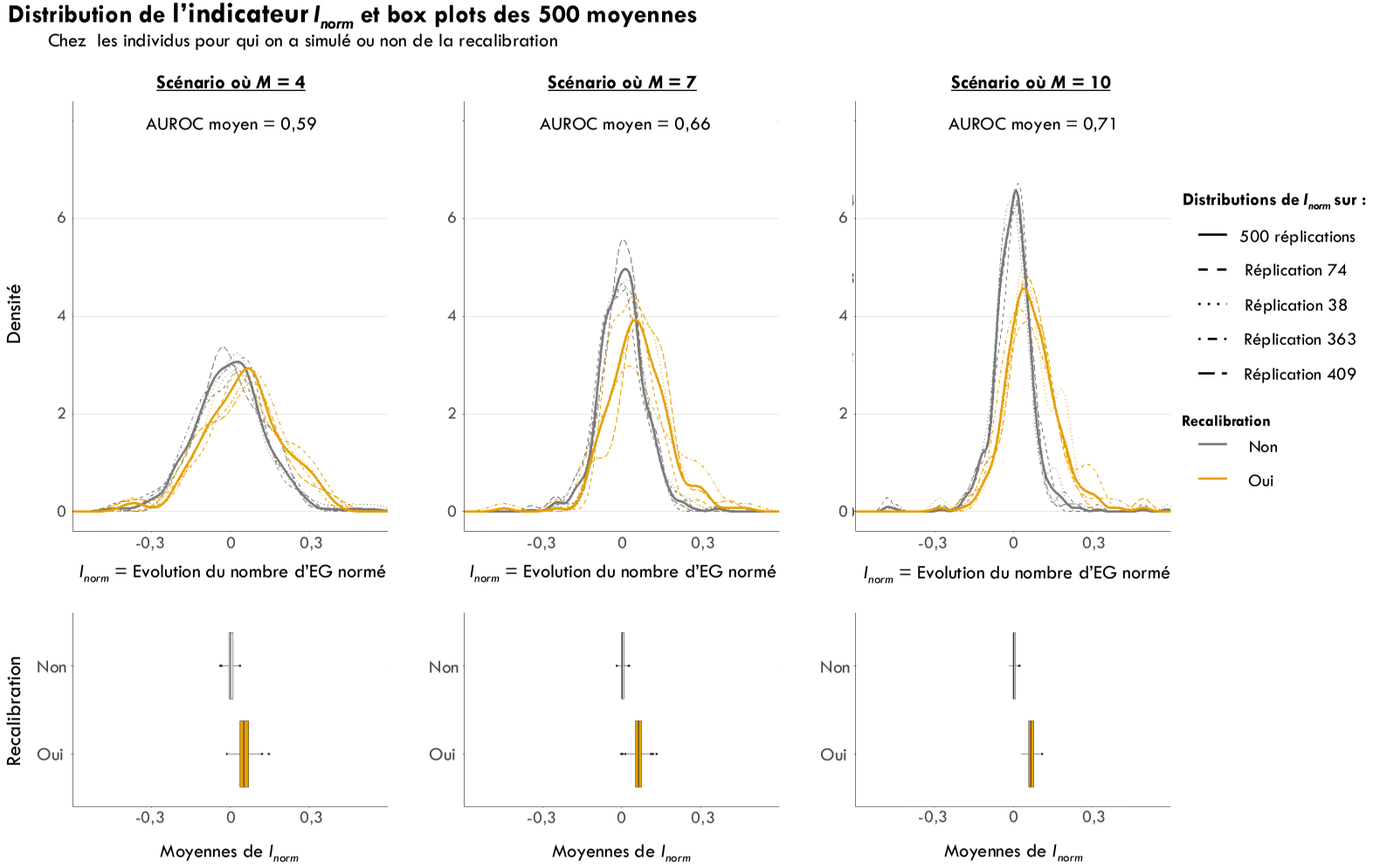


FIGURE 4.7 – *Box plots* des moyennes de **l'indicateur I_{norm}** chez les patients avec et sans recalibration et distribution de l'indicateur dans ces mêmes sous-groupes. Trois scénarios considérés : N = 200 individus, J = 7 items, M = 4, 7 ou 10 modalités de réponse, Recalibration uniforme simulée pour 25% de l'échantillon et touchant J_{RS} = 2 items, Moyenne de la variable latente qui se détériore en moyenne ($\Delta = -0,2$)

4.3.3 Hypothèse soulevée par cette 1re étude de simulation

En s'intéressant aux paramètres de seuil des items du modèle que nous avons utilisé pour simuler les données, nous avons observé que la recalibration non uniforme n'avait parfois que peu modifié la façon dont ces paramètres s'ordonnaient les uns par rapport aux autres (comparé à la recalibration uniforme). Pour simplifier, nous désignerons ces paramètres de seuil par le terme "paramètres d'item" dans les prochains paragraphes.

Afin d'illustrer cette observation, considérons les scénarios avec $J = 7$ items, $M = 4$ modalités de réponse et où la recalibration (uniforme ou non uniforme) se manifeste sur $J_{RS} = 2$ items. Les paramètres des items du LPCM pour ces scénarios (notés $\delta_{jp}^{(t)}$) ont été regroupés dans la figure 4.8. Pour chaque bloc de cette figure, un rang a été affecté à ces paramètres d'item, afin de quantifier l'effet de la recalibration sur la façon dont ils s'ordonnent les uns par rapport aux autres à chaque temps.

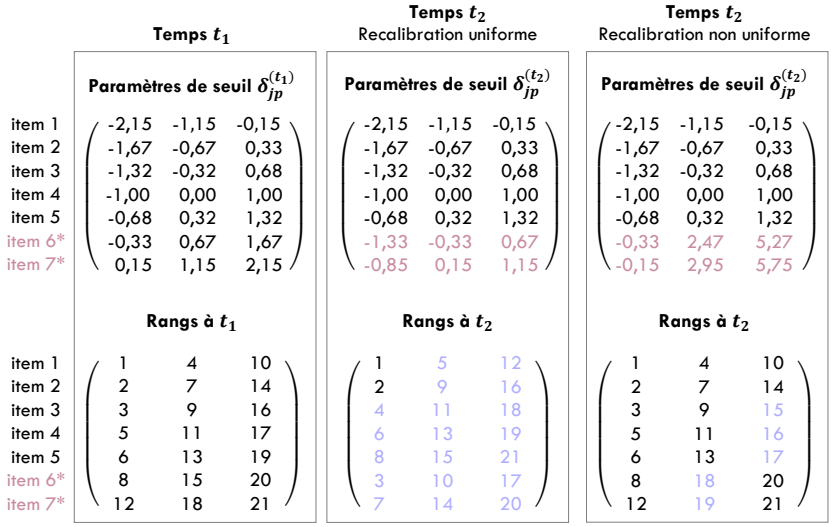


FIGURE 4.8 – Paramètres de seuil des items $\delta_{jp}^{(t)}$ utilisés pour la génération des données des individus avec recalibration simulée (1re étude de simulation). Ces paramètres sont accompagnés de leur rang à chaque temps.

Notes : *Les items sur lesquels la recalibration se manifeste sont indiqués en rose avec un astérisque. Les rangs impactés par la recalibration sont de couleur mauve.*

Pour les scénarios considérés dans cet exemple, et lorsque la recalibration est uniforme, il y a un écart absolu d'en moyenne 2,2 entre les rangs des paramètres d'item au temps t_1 et les rangs des paramètres d'item au temps t_2. Cet écart absolu moyen n'est que de 0,4 lorsque la recalibration est non uniforme.

Cette observation s'est avérée valable pour tous les scénarios considérés dans l'étude de simulation : à paramètres de simulation identiques, la recalibration non uniforme impactait systématiquement moins l'ordre des paramètres d'item que la recalibration uniforme. De fait, les paramètres d'item que nous avons décalés pour opérationnaliser la recalibration se trouvaient, au temps t_1, à une position relativement élevée sur le continuum de la variable latente (voir figure 4.9 pour une illustration avec $J = 7$ items).

Avec la recalibration uniforme, tous les paramètres des items touchés par la recalibration ont été décalés vers le bas, entraînant des modifications dans l'ordre des paramètres d'item entre le temps t_1 et le temps t_2.

En revanche, avec la recalibration non uniforme, ces paramètres ont été décalés vers le haut (alors qu'ils étaient déjà en position élevée), n'entraînant ainsi que peu de modifications dans l'ordre des paramètres d'item entre les deux temps de mesure. Un autre effet vient s'ajouter à celui de la "direction du décalage" : lorsque la recalibration était non uniforme, le premier paramètre des items touchés par la recalibration (δ_{j1}) restait fixe entre les deux temps de mesure.

Le faible impact de la recalibration non uniforme sur la façon dont s'ordonne les paramètres d'item est très probablement à l'origine des moins bonnes performances des indicateurs parmi les scénarios avec cette forme de recalibration. Un exemple de ces moins bons résultats est donné dans la figure 4.10 pour les deux indicateurs.

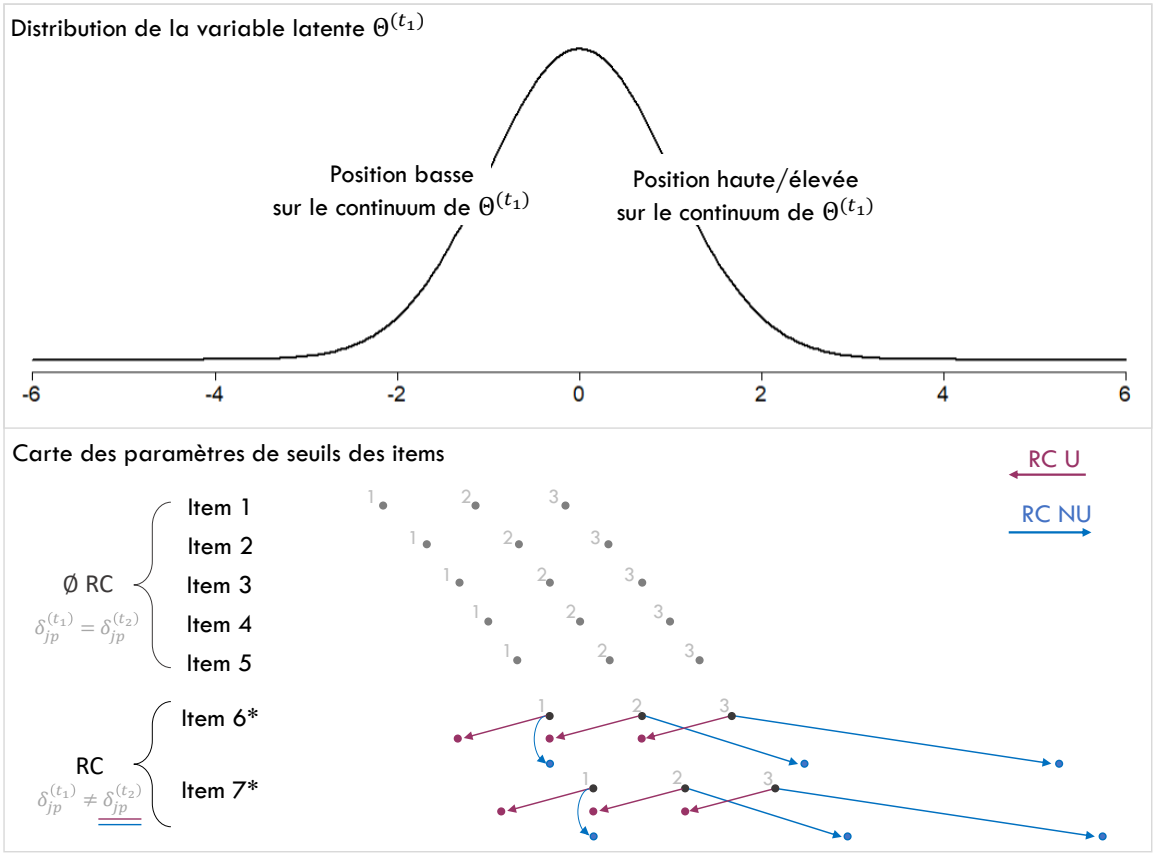


FIGURE 4.9 – Direction de la recalibration uniforme (RC U) et non uniforme (RC NU) lors de la première étude de simulation (scénarios avec $J = 7$ items, $M = 4$ modalités de réponses et où la recalibration se manifeste sur $J_{RS} = 2$ items)

Notes :

Les points représentent les paramètres de seuil des items utilisés pour la simulation des données. Ils sont représentés le long du continuum de la variable latente au temps t_1. La recalibration (RC) se manifeste sur les items 6 et 7 : leurs paramètres de seuil sont décalés au deuxième temps de mesure. Ils sont soit décalés vers la gauche (pour la recalibration uniforme) soit vers la droite (pour la recalibration non uniforme). Ces décalages sont matérialisés par les flèches. Les items 1 à 5 ne sont pas touchés par la recalibration : leurs paramètres de seuil restent donc constants.

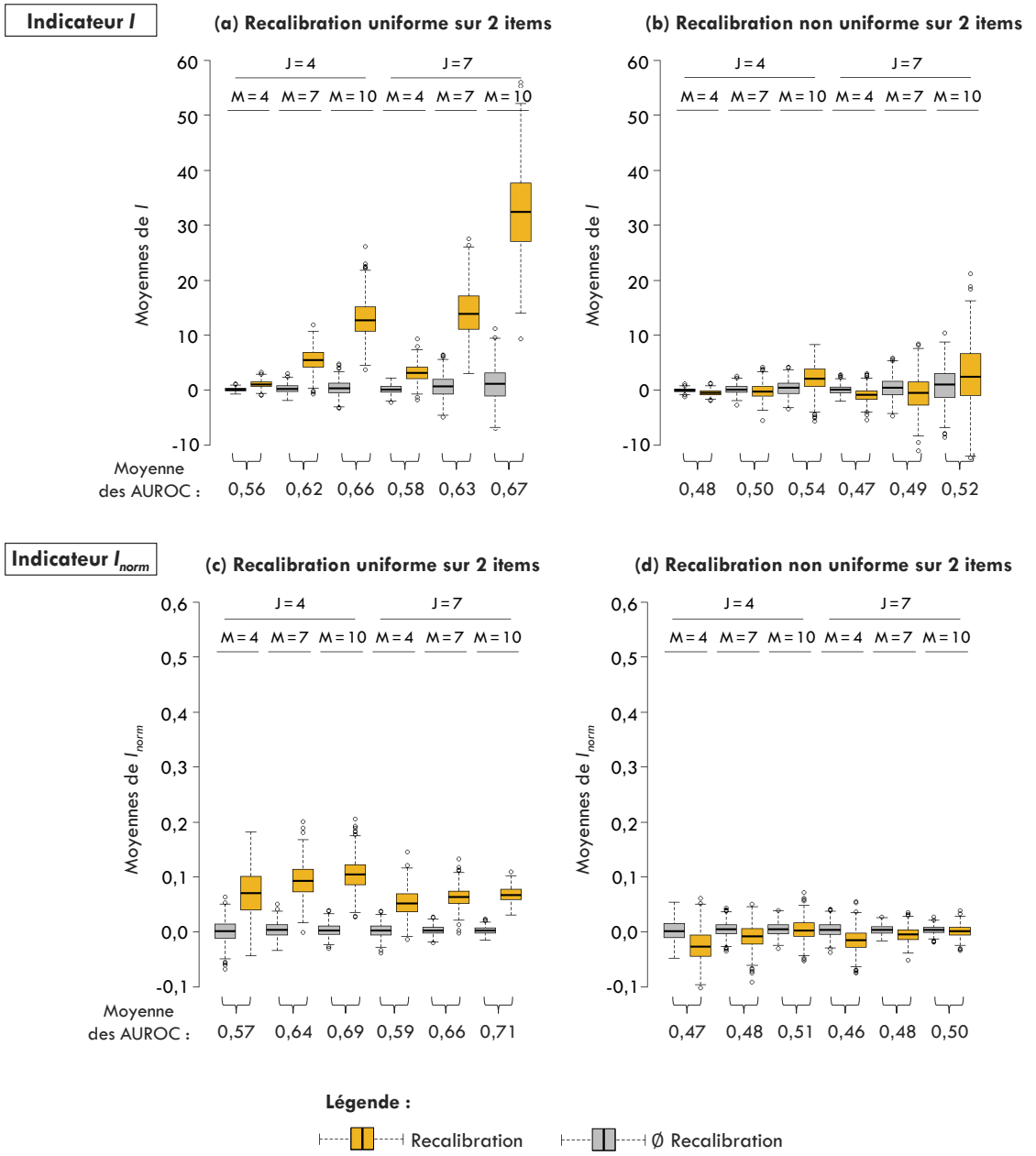


FIGURE 4.10 – *Box plots* des 500 moyennes des indicateurs I et I_{norm} obtenues séparément chez les individus avec et sans recalibration simulée. Chaque couple de *box plot* correspond à un scénario. Les aires sous la courbe ROC (AUROC) moyennes pour chacun des scénarios sont indiquées par les accolades. Scénarios considérés : N = 200 individus, J = 4 ou 7 items, M = 4, 7 ou 10 modalités de réponse, Δ = −0,2 (variable latente qui se détériore en moyenne), p = 25% des individus expérimentent la recalibration simulée (uniforme ou non uniforme), Recalibration qui se manifeste sur J_{RS} = 2 items.

Suite à ces résultats, l'hypothèse suivante a été soulevée : la position des items sur lesquels se manifeste la recalibration aurait un impact sur les performances des indicateurs I et I_{norm}. Cette hypothèse a été explorée dans une seconde étude de simulation présentée dans la section suivante.

4.4 2e étude de simulation

4.4.1 Plan de simulation et analyse

Cette étude de simulation est inspirée de la première qui a été présentée dans la section 4.3. Un nouveau paramètre de simulation a été exploré : la position initiale des items touchés par la recalibration. Le plan de simulation est détaillé ci-dessous.

Pour cette étude de simulation, nous avons simulé les réponses de $N = 200$ individus à un questionnaire unidimensionnel composé de $J = 4$ ou 7 items polytomiques (item 1, item 2, ..., item J) et complété à deux temps de mesure t_1 et t_2. Il y avait pour chaque item $M = 4$, 7 ou 10 modalités de réponse (allant de 0 à $M-1$), et nous avons simulé du *response shift* (recalibration uniquement) pour une partie des individus ($p = 25\%$). Les niveaux individuels de la variable latente aux deux temps de mesure ont été tirés dans une loi binormale :

$$\boldsymbol{\Theta} = \begin{pmatrix} \Theta^{(t_1)} \\ \Theta^{(t_2)} \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu^{(t_1)} \\ \mu^{(t_2)} \end{bmatrix}, \Sigma\right)$$

Avec :

- $\Theta^{(t_1)}$ et $\Theta^{(t_2)}$ les variables latentes aux temps de mesure t_1 et t_2 ;
- $\mu^{(t_1)}$ et $\mu^{(t_2)}$ les moyennes respectives de $\Theta^{(t_1)}$ et $\Theta^{(t_2)}$. Trois valeurs ont été considérées pour l'évolution du niveau moyen de la variable latente ($\Delta = \mu^{(t_2)} - \mu^{(t_1)}$) : $\Delta = -0{,}2,\ 0$ ou $0{,}2$;
- Σ la matrice de variance-covariance des variables latentes, elle est définie par :

$$\Sigma = \begin{bmatrix} 1 & 0{,}6 \\ 0{,}6 & 1 \end{bmatrix}$$

Les réponses des individus ont été générées selon un LPCM dont la formulation est rappelée par l'équation 4.1 :

$$P\left(X_{ij}^{(t)} = x \mid \theta_i^{(t)}, \delta_{j1}^{(t)}, \ldots, \delta_{jM-1}^{(t)}\right) = \frac{\exp\left(x\theta_i^{(t)} - \sum_{p=1}^{x} \delta_{jp}^{(t)}\right)}{\sum_{l=0}^{M-1} \exp\left(l\theta_i^{(t)} - \sum_{p=1}^{l} \delta_{jp}^{(t)}\right)} \tag{4.1}$$

Où :

- $X_{ij}^{(t)}$ désigne la réponse à l'item $j = 1, \ldots, J$ de l'individu $i = 1, \ldots, N$ au temps $t = t_1$ ou t_2.
- $\theta_i^{(t)}$ représente le niveau de variable latente de l'individu i au temps t (réalisation de la variable latente $\Theta^{(t)}$).
- Les $\delta_{jp}^{(t)}$ désignent les paramètres de seuil de l'item j au temps t.

  Il y a un paramètre de seuil pour chaque modalité de réponse p supérieure à 0 ($p = 1, ..., M-1$). Ces paramètres ont été choisis pour (i) être centrés sur 0 (la moyenne de la variable latente au premier temps de mesure) et (ii) couvrir tout le continuum de $\Theta^{(t_1)}$.

Les paramètres de seuil des items au temps t_1 sont les mêmes que ceux de la première étude de simulation. Pour rappel, ils dérivent des quantiles d'une loi normale centrée-réduite (cf. l'annexe de l'article page 129). Les valeurs de ces paramètres sont regroupées dans le tableau 4.2.

TABLEAU 4.2 – Paramètres de seuil des items $\delta_{jp}^{(t_1)}$ utilisés pour l'étude de simulation

	M = 4			**M = 7**						**M = 10**								
	$\delta_{j1}^{(t_1)}$	$\delta_{j2}^{(t_1)}$	$\delta_{j3}^{(t_1)}$	$\delta_{j1}^{(t_1)}$	$\delta_{j2}^{(t_1)}$	$\delta_{j3}^{(t_1)}$	$\delta_{j4}^{(t_1)}$	$\delta_{j5}^{(t_1)}$	$\delta_{j6}^{(t_1)}$	$\delta_{j1}^{(t_1)}$	$\delta_{j2}^{(t_1)}$	$\delta_{j3}^{(t_1)}$	$\delta_{j4}^{(t_1)}$	$\delta_{j5}^{(t_1)}$	$\delta_{j6}^{(t_1)}$	$\delta_{j7}^{(t_1)}$	$\delta_{j8}^{(t_1)}$	$\delta_{j9}^{(t_1)}$
***J* = 4**																		
item 1	-1,84	-0,84	0,16	-1,84	-1,44	-1,04	-0,64	-0,24	0,16	-1,84	-1,59	-1,34	-1,09	-0,84	-0,59	-0,34	-0,09	0,16
item 2	-1,25	-0,25	0,75	-1,25	-0,85	-0,45	-0,05	0,35	0,75	-1,25	-1,00	-0,75	-0,50	-0,25	0,00	0,25	0,50	0,75
item 3	-0,75	0,25	1,25	-0,75	-0,35	0,05	0,45	0,85	1,25	-0,75	-0,50	-0,25	0,00	0,25	0,50	0,75	1,00	1,25
item 4	-0,16	0,84	1,84	-0,16	0,24	0,64	1,04	1,44	1,84	-0,16	0,09	0,34	0,59	0,84	1,09	1,34	1,59	1,84
***J* = 7**																		
item 1	-2,15	-1,15	-0,15	-2,15	-1,75	-1,35	-0,95	-0,55	-0,15	-2,15	-1,90	-1,65	-1,40	-1,15	-0,90	-0,65	-0,40	-0,15
item 2	-1,67	-0,67	0,33	-1,67	-1,27	-0,87	-0,47	-0,07	0,33	-1,67	-1,42	-1,17	-0,92	-0,67	-0,42	-0,17	0,08	0,33
item 3	-1,32	-0,32	0,68	-1,32	-0,92	-0,52	-0,12	0,28	0,68	-1,32	-1,07	-0,82	-0,57	-0,32	-0,07	0,18	0,43	0,68
item 4	-1,00	0,00	1,00	-1,00	-0,60	-0,20	0,20	0,60	1,00	-1,00	-0,75	-0,50	-0,25	0,00	0,25	0,50	0,75	1,00
item 5	-0,68	0,32	1,32	-0,68	-0,28	0,12	0,52	0,92	1,32	-0,68	-0,43	-0,18	0,07	0,32	0,57	0,82	1,07	1,32
item 6	-0,33	0,67	1,67	-0,33	0,07	0,47	0,87	1,27	1,67	-0,33	-0,08	0,17	0,42	0,67	0,92	1,17	1,42	1,67
item 7	0,15	1,15	2,15	0,15	0,55	0,95	1,35	1,75	2,15	0,15	0,40	0,65	0,90	1,15	1,40	1,65	1,90	2,15

Notes :

J : Nombre d'items

M : Nombre de modalités de réponse par item

$\delta_{jp}^{(t_1)}$ *: Paramètre de seuil associé à la modalité de réponse p de l'item j au temps* t_1

Opérationnalisation de la recalibration

Comme dans l'étude de simulation précédente, la recalibration n'a été simulée que sur une partie des individus. Chez ces individus, elle a été opérationnalisée en faisant varier, au cours du temps, les paramètres de seuil de certains items (pour les individus sans recalibration simulée, ces paramètres restaient constants). Au sein d'un même scénario, on a simulé des individus qui expérimenteraient tous exactement la même recalibration : même forme, même taille (ou magnitude) et mêmes items touchés. Ainsi, au temps t_2, les paramètres de seuil des items touchés par la recalibration sont donnés par l'équation 4.2, où η_{jp} est le paramètre modélisant la recalibration :

$$\forall\ p \in \{1, \ldots, M-1\},\quad \delta_{jp}^{(t_2)} = \begin{cases} \delta_{jp}^{(t_1)} + \eta_{jp} & \text{pour les individus avec recalibration} \\ \delta_{jp}^{(t_1)} & \text{pour les individus sans recalibration} \end{cases} \tag{4.2}$$

Deux formes de recalibration ont été étudiées :

- **La recalibration uniforme :** les paramètres de seuil d'un item j touché par de la recalibration uniforme ont tous été décalés dans la même direction et avec la même magnitude. Pour l'étude de simulation, nous avons utilisé : $\forall\ p \in \{1, \ldots, M-1\},\quad \eta_{jp} = \eta = +1$.

- **La recalibration non uniforme :** les paramètres de seuil d'un item touché par de la recalibration non uniforme sont décalés, mais la direction et/ou la magnitude de ces décalages varient.

  Cette étude de simulation est restreinte au cas où les décalages varient en magnitude, mais conservent la même direction. Les différentes magnitudes des décalages sont données par la formule 4.3 :

$$\eta_{jp} = \begin{cases} \dfrac{(p-1)\eta}{p} & \text{si } 1 \leq p < M/2 \\ \eta & \text{si } p = M/2 \\ \dfrac{(M-p+1)\eta}{M-p} & \text{si } M/2 < p \leq M-1 \end{cases} \qquad \text{avec } \eta = +1 \tag{4.3}$$

Ces deux formes de recalibration sont représentées graphiquement dans la figure 4.11.

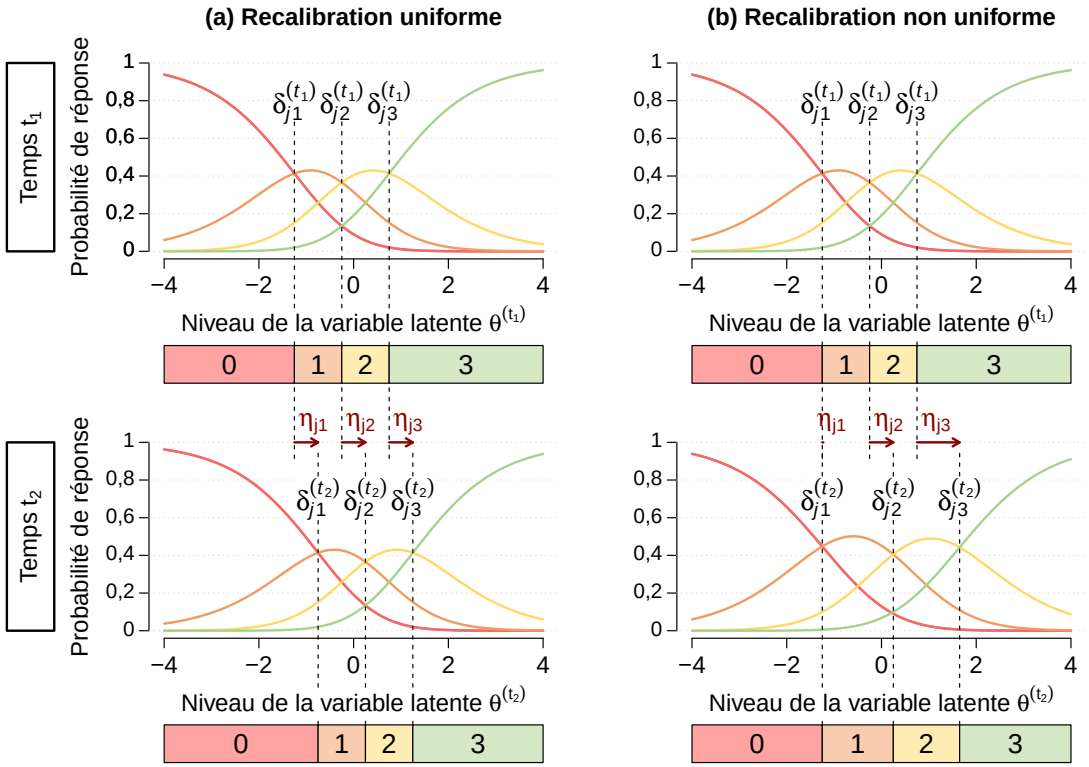


FIGURE 4.11 – Courbes caractéristiques des modalités de réponse d'un item j touché par de la recalibration uniforme (graphique a) ou non uniforme (graphique b)

Notes :
Le décalage dans le temps des paramètres de seuil des items est représenté par les flèches rouges.
Lorsque la recalibration est uniforme, ces flèches vont dans la même direction et ont la même taille. Lorsque la recalibration est non uniforme, nous avons choisi de faire uniquement varier la taille des flèches (elles conservent la même direction).

Contrairement à la première étude de simulation, la direction des deux formes de recalibration a été harmonisée (les paramètres de seuil des items touchés par la recalibration sont dorénavant décalés vers la droite en leur ajoutant des quantités positives). Pour la recalibration non uniforme, on peut noter que le premier paramètre $\delta_{j1}^{(t)}$ reste constant au cours du temps. Les paramètres suivants sont en revanche décalés vers la droite avec une magnitude croissante. Pour chaque item touché par la recalibration, la moyenne des paramètres η_{jp} était égale à $+1$ (comme pour la recalibration uniforme).

La recalibration a été simulée selon les règles suivantes :

- Elle pouvait se manifester sur :
  - J_{RS} = 1 ou 2 items pour les scénarios où $J = 4$;
  - J_{RS} = 1, 2 ou 3 items pour les scénarios où $J = 7$.
- Les items touchés par la recalibration étaient les mêmes pour tous les individus pour qui on simulait de la recalibration.
- Au sein d'un jeu de données, la recalibration était la même sur tous les items (même forme et même taille).

Les items touchés par la recalibration ont été sélectionnés dans l'objectif d'explorer l'impact potentiel de leur position sur les performances des indicateurs I et I_{norm}. Ces items pouvaient donc être en position :

- **Moyenne :** Au temps t_1, les paramètres de seuil des items touchés par la recalibration sont autour de $\mu^{(t_1)} = 0$ (la moyenne de la variable latente $\Theta^{(t_1)}$).
- **Basse :** Au temps t_1, les paramètres de seuil des items touchés par la recalibration sont dans la partie basse de la distribution de la variable latente $\Theta^{(t_1)}$.
- **Haute :** Au temps t_1, les paramètres de seuil des items touchés par la recalibration sont dans la partie haute de la distribution de la variable latente $\Theta^{(t_1)}$.
- **Extrême :** Au temps t_1, les paramètres de seuil des items touchés par la recalibration sont à la fois dans la partie basse et la partie haute de la distribution de la variable latente $\Theta^{(t_1)}$.

L'identification des items touchés par la recalibration pour chacune de ces configurations est donnée dans le tableau 4.3 en fonction de J et J_{RS}.

TABLEAU 4.3 – Identification des items touchés par la recalibration pour les différentes positions explorées : Moyenne (Moy.), Basse, Haute et Extrême (Extr.)

	J_{RS} = 1				J_{RS} = 2			
J = 4	Moy.	Basse	Haute	Extr.	Moy.	Basse	Haute	Extr.
Item 1		✗				✗		✗
Item 2	✗				✗	✗		
Item 3					✗		✗	
Item 4			✗				✗	✗

	J_{RS} = 1				J_{RS} = 2				J_{RS} = 3			
J = 7	Moy.	Basse	Haute	Extr.	Moy.	Basse	Haute	Extr.	Moy.	Basse	Haute	Extr.
Item 1		✗				✗		✗		✗		✗
Item 2						✗				✗		✗
Item 3					✗				✗	✗		
Item 4	✗								✗			
Item 5					✗				✗		✗	
Item 6							✗				✗	
Item 7			✗				✗	✗			✗	✗

Notes :
Les items sur lesquels se manifeste la recalibration sont identifiés par les croix ✗
J : Nombre d'items dans le questionnaire
J_{RS} : Nombre d'items touchés par la recalibration

L'étude de simulation est résumée dans le tableau 4.4. La combinaison de tous les paramètres de simulation a conduit à un total de 324 scénarios, tous répliqués 500 fois.

Analyse des jeux de données

Pour chaque jeu de données, les réalisations des indicateurs I et I_{norm} ont été calculées pour chaque individu, puis la capacité discriminante de ces indicateurs a été évaluée en estimant les AUROC associées. Les 500 AUROC obtenues pour chaque scénario et chaque indicateur ont ensuite été résumées par leur moyenne et leur écart-type.

TABLEAU 4.4 – Résumé de la deuxième étude de simulation

Échantillon et structure du questionnaire	
Taille de l'échantillon	$N = 200$ individus simulés
Nombre d'items	$J = 4, 7$ items
Nombre de modalités de réponse	$M = 4$; 7 ; 10 modalités de réponse
Variables latentes $\Theta^{(t_1)}$ et $\Theta^{(t_2)}$	
Moyenne au temps t_1	$\mu^{(t_1)} = 0$
Moyenne au temps t_2	$\mu^{(t_2)} = \mu^{(t_1)} + \Delta$ avec $\Delta = -0{,}2\,;\ 0\,;\ 0,2$
Ecart-type au temps t_1	$\sigma^{(t_1)} = 1$
Ecart-type au temps t_2	$\sigma^{(t_2)} = 1$
Covariance entre $\Theta^{(t_1)}$ et $\Theta^{(t_2)}$	0,6
Recalibration	
Proportion d'individus concernés	$p = 25\%$
Nombre d'items touchés	$J_{RS} = 1$; 2 si $J = 4$
	$J_{RS} = 1$; 2 ; 3 si $J = 7$
Forme de la recalibration	
Uniforme :	Les paramètres de seuil d'un item touché par la recalibration uniforme sont tous décalés dans la même direction et avec la même magnitude
Non uniforme :	Les décalages varient en magnitude, mais conservent la même direction
Taille de la recalibration	Les paramètres de seuil d'un item touché par la recalibration augmentent en moyenne de $\eta = +1$
Position à t_1 des items touchés	
Moyenne :	Les paramètres de seuil des items touchés par la recalibration sont initialement centrés sur $\mu^{(t_1)} = 0$
Basse :	Les paramètres de seuil des items touchés par la recalibration sont initialement dans la partie basse de la distribution de $\Theta^{(t_1)}$
Haute :	Les paramètres de seuil des items touchés par la recalibration sont initialement dans la partie haute de la distribution de $\Theta^{(t_1)}$
Extrême :	Les paramètres de seuil des items touchés par la recalibration sont initialement dans les parties basse et haute de la distribution de $\Theta^{(t_1)}$

Outils logiciels

Ces travaux ont été réalisés à l'aide du logiciel Stata version 15 (StataCorp, College station, TX). La simulation des données a été réalisée grâce au module Stata *simirt* [163] qui permet de générer des données PRO selon différents modèles de la famille de Rasch ou de la famille de Lord (items dichotomiques et polytomiques). Le module Stata *eguttman* [2], que j'ai développé lors de la première étude de simulation, a été réutilisé pour cette seconde étude. Pour rappel, ce module est disponible depuis la plate-forme OSF (lien cliquable) et est succinctement présentée an page 137.

4.4.2 Résultats

Les moyennes et les écart-types des 500 AUROC obtenues pour les indicateurs I et I_{norm} sont résumées dans les tableaux 4.5, 4.6 et 4.7 pour les scénarios où Δ (l'évolution moyenne de la variable latente) vaut -0,2 , 0 et 0,2, respectivement.

Lorsque les items touchés par la recalibration étaient en position "haute", les moyennes des AUROC associées aux indicateurs I et I_{norm} n'excédaient jamais 0,57 et 0,53, respectivement. Cela signifie que dans ces scénarios, les indicateurs ne permettent pas de discriminer les individus pour lesquels on a simulé de la recalibration des autres (pour lesquels on n'en a pas simulé). Il s'agit là des pires performances observées pour les deux indicateurs. Ces résultats étaient attendus, car lorsque les items touchés par la recalibration sont en position "haute", l'ordre des paramètres de seuil des items n'est quasiment pas impacté entre t_1 et t_2. Les AUROC moyennes étaient en revanche systématiquement supérieures lorsque la position des items touchés par la recalibration était "moyenne", "basse" ou "extrême" (comparé à celles observées avec la position "haute"). Les AUROC moyennes les plus hautes ont été obtenues lorsque les items touchés par la recalibration étaient en position "basse".

2. Le module *eguttman* permet de compter le nombre d'erreurs de Guttman, déterminer le nombre maximal d'erreurs de Guttman possiblement atteignable et calculer le nombre d'erreurs de Guttman normé pour chaque individu d'une base de données en utilisant un ordre de difficulté observé à un certain temps de mesure.

De plus, contrairement à la première étude de simulation, les AUROC moyennes obtenues parmi les scénarios avec de la recalibration uniforme sont globalement similaires à celles obtenues parmi les scénarios avec de la recalibration non uniforme.

L'ensemble de ces résultats permet de confirmer l'hypothèse que nous avions émise : la position des items touchés par la recalibration influence effectivement les performances de nos deux indicateurs. Ces résultats sont néanmoins à nuancer, puisque la capacité discriminante de nos deux indicateurs n'est pas satisfaisante. En effet, si l'on exclut les scénarios avec $M = 10$ modalités de réponse (cas exploré pour approcher une échelle visuelle analogique), les AUROC moyennes ne dépassent pas 0,66 pour l'indicateur I et 0,69 pour l'indicateur I_{norm}. Cela signifie que les indicateurs ne permettent pas de distinguer les individus pour lesquels on a simulé de la recalibration des autres individus pour lesquels on n'en a pas simulé (les réalisations des indicateurs chez les individus avec recalibration restant trop proches des réalisations des indicateurs chez les individus sans recalibration).

TABLEAU 4.5 – Moyennes des 500 aires sous la courbe ROC (AUROC) associées aux indicateurs I et I_{norm} en fonction de la forme de la recalibration (forme RC), du nombre d'items touchés par la recalibration (J_{RS}), du nombre total d'items (J), du nombre de modalités de réponse (M) et de la position des items touchés par la recalibration (Moyenne, Basse, Haute ou Extrême).
Les scénarios considérés dans ce tableau sont ceux où la variable latente diminue en moyenne entre t_1 et t_2 (c.-à-d. Δ = -0,2)

$\Delta = -0,2$

Forme RC	J_{RS}	J	M	**Moyenne des AUROC associées à l'évolution du du nombre d'EG I** Position : Moy.	Basse	Haute	Extr.	**Moyenne des AUROC associées à l'évolution du du nombre d'EG normé I_{norm}** Position : Moy.	Basse	Haute	Extr.
U	1	4	4	0,52	0,55	0,50		0,52	0,56	0,49	
U	1	4	7	0,57	0,59	0,51		0,57	0,62	0,50	
U	1	4	10	0,60	0,63	0,53		0,60	0,66	0,51	
U	1	7	4	0,52	0,54	0,50		0,52	0,55	0,49	
U	1	7	7	0,55	0,58	0,50		0,55	0,61	0,49	
U	1	7	10	0,58	0,61	0,51		0,57	0,64	0,50	
U	2	4	4	0,52	0,55	0,50	0,52	0,52	0,57	0,49	0,54
U	2	4	7	0,56	0,59	0,51	0,57	0,55	0,64	0,50	0,59
U	2	4	10	0,59	0,63	0,55	0,61	0,59	0,68	0,52	0,64
U	2	7	4	0,54	0,57	0,49	0,53	0,54	0,60	0,48	0,54
U	2	7	7	0,57	0,62	0,50	0,57	0,57	0,67	0,48	0,59
U	2	7	10	0,61	0,66	0,52	0,61	0,61	0,72	0,50	0,63
U	3	7	4	0,53	0,57	0,48	0,55	0,53	0,60	0,47	0,57
U	3	7	7	0,57	0,62	0,50	0,60	0,57	0,68	0,48	0,64
U	3	7	10	0,61	0,67	0,54	0,64	0,62	0,74	0,51	0,70
NU	1	4	4	0,51	0,53	0,49		0,50	0,53	0,48	
NU	1	4	7	0,54	0,59	0,49		0,54	0,59	0,48	
NU	1	4	10	0,58	0,62	0,51		0,57	0,63	0,49	
NU	1	7	4	0,50	0,54	0,49		0,50	0,55	0,49	
NU	1	7	7	0,52	0,58	0,50		0,52	0,60	0,49	
NU	1	7	10	0,55	0,62	0,50		0,54	0,64	0,49	
NU	2	4	4	0,49	0,51	0,48	0,51	0,48	0,52	0,47	0,51
NU	2	4	7	0,53	0,58	0,49	0,57	0,52	0,59	0,47	0,56
NU	2	4	10	0,56	0,62	0,51	0,60	0,55	0,64	0,48	0,60
NU	2	7	4	0,50	0,54	0,48	0,53	0,49	0,55	0,47	0,53
NU	2	7	7	0,54	0,61	0,48	0,57	0,53	0,63	0,47	0,58
NU	2	7	10	0,57	0,66	0,50	0,62	0,57	0,69	0,48	0,63
NU	3	7	4	0,48	0,54	0,46	0,53	0,48	0,54	0,45	0,53
NU	3	7	7	0,53	0,61	0,48	0,60	0,52	0,63	0,47	0,61
NU	3	7	10	0,57	0,67	0,50	0,65	0,56	0,70	0,47	0,67

Notes
U : Uniforme, NU : Non uniforme
Pour l'ensemble des scénarios de ce tableau, les écart-types des 500 AUROC oscillaient entre 0,04 et 0,05.

TABLEAU 4.6 – Moyennes des 500 aires sous la courbe ROC (AUROC) associées aux indicateurs I et I_{norm} en fonction de la forme de la recalibration (forme RC), du nombre d'items touchés par la recalibration (J_{RS}), du nombre total d'items (J), du nombre de modalités de réponse (M) et de la position des items touchés par la recalibration (Moyenne, Basse, Haute ou Extrême).
Les scénarios considérés dans ce tableau sont ceux où la variable latente reste stable en moyenne entre t_1 et t_2 (c.-à-d. $\Delta = 0$)

$\Delta = 0$				**Moyenne des AUROC associées à l'évolution du du nombre d'EG I**				**Moyenne des AUROC associées à l'évolution du du nombre d'EG normé I_{norm}**			
				Position :				Position :			
Forme RC	J_{RS}	J	M	Moy.	Basse	Haute	Extr.	Moy.	Basse	Haute	Extr.
U	1	4	4	0,53	0,56	0,50		0,53	0,56	0,49	
U	1	4	7	0,58	0,60	0,52		0,57	0,62	0,50	
U	1	4	10	0,62	0,63	0,54		0,61	0,66	0,51	
U	1	7	4	0,52	0,54	0,50		0,52	0,55	0,49	
U	1	7	7	0,56	0,58	0,50		0,55	0,60	0,49	
U	1	7	10	0,59	0,61	0,52		0,58	0,64	0,51	
U	2	4	4	0,53	0,55	0,50	0,53	0,53	0,57	0,49	0,54
U	2	4	7	0,56	0,61	0,52	0,58	0,56	0,64	0,50	0,59
U	2	4	10	0,61	0,65	0,56	0,63	0,60	0,69	0,52	0,63
U	2	7	4	0,54	0,57	0,49	0,54	0,53	0,59	0,48	0,54
U	2	7	7	0,58	0,63	0,52	0,57	0,58	0,67	0,50	0,58
U	2	7	10	0,63	0,67	0,53	0,61	0,62	0,72	0,51	0,63
U	3	7	4	0,54	0,58	0,49	0,55	0,54	0,61	0,47	0,57
U	3	7	7	0,59	0,64	0,52	0,61	0,59	0,69	0,49	0,64
U	3	7	10	0,63	0,69	0,55	0,67	0,63	0,75	0,52	0,70
NU	1	4	4	0,51	0,54	0,49		0,51	0,54	0,48	
NU	1	4	7	0,55	0,60	0,49		0,54	0,60	0,48	
NU	1	4	10	0,60	0,64	0,51		0,58	0,65	0,49	
NU	1	7	4	0,50	0,54	0,49		0,50	0,54	0,48	
NU	1	7	7	0,53	0,60	0,49		0,52	0,61	0,48	
NU	1	7	10	0,56	0,63	0,50		0,55	0,64	0,49	
NU	2	4	4	0,49	0,53	0,48	0,52	0,48	0,53	0,47	0,51
NU	2	4	7	0,54	0,60	0,49	0,58	0,52	0,60	0,47	0,57
NU	2	4	10	0,57	0,64	0,52	0,62	0,56	0,65	0,49	0,61
NU	2	7	4	0,50	0,56	0,48	0,53	0,49	0,56	0,47	0,53
NU	2	7	7	0,55	0,63	0,49	0,58	0,53	0,64	0,47	0,58
NU	2	7	10	0,59	0,69	0,50	0,63	0,57	0,71	0,48	0,63
NU	3	7	4	0,49	0,55	0,46	0,55	0,48	0,55	0,45	0,54
NU	3	7	7	0,54	0,63	0,48	0,62	0,52	0,64	0,46	0,62
NU	3	7	10	0,59	0,69	0,50	0,68	0,57	0,71	0,47	0,69

Notes
U : Uniforme, NU : Non uniforme
Pour l'ensemble des scénarios de ce tableau, les écart-types des 500 AUROC oscillaient entre 0,04 et 0,05.

TABLEAU 4.7 – Moyennes des 500 aires sous la courbe ROC (AUROC) associées aux indicateurs I et I_{norm} en fonction de la forme de la recalibration (forme RC), du nombre d'items touchés par la recalibration (J_{RS}), du nombre total d'items (J), du nombre de modalités de réponse (M) et de la position des items touchés par la recalibration (Moyenne, Basse, Haute ou Extrême).
Les scénarios considérés dans ce tableau sont ceux où la variable latente augmente en moyenne entre t_1 et t_2 (c.-à-d. $\Delta = 0{,}2$)

$\Delta = 0,2$				**Moyenne des AUROC associées à l'évolution du du nombre d'EG I**				**Moyenne des AUROC associées à l'évolution du du nombre d'EG normé I_{norm}**			
				Position :				Position :			
Forme RC	J_{RS}	J	M	Moy.	Basse	Haute	Extr.	Moy.	Basse	Haute	Extr.
U	1	4	4	0,54	0,55	0,51		0,53	0,56	0,49	
U	1	4	7	0,58	0,60	0,52		0,58	0,61	0,5	
U	1	4	10	0,62	0,63	0,55		0,62	0,65	0,51	
U	1	7	4	0,53	0,55	0,50		0,53	0,56	0,49	
U	1	7	7	0,56	0,58	0,51		0,55	0,60	0,50	
U	1	7	10	0,60	0,61	0,53		0,59	0,63	0,51	
U	2	4	4	0,54	0,56	0,51	0,54	0,53	0,57	0,5	0,54
U	2	4	7	0,58	0,62	0,54	0,59	0,57	0,64	0,51	0,59
U	2	4	10	0,63	0,66	0,57	0,64	0,61	0,69	0,53	0,63
U	2	7	4	0,55	0,57	0,49	0,54	0,54	0,59	0,48	0,54
U	2	7	7	0,59	0,64	0,52	0,58	0,58	0,67	0,49	0,58
U	2	7	10	0,64	0,68	0,56	0,62	0,63	0,71	0,52	0,62
U	3	7	4	0,55	0,59	0,50	0,56	0,54	0,61	0,48	0,57
U	3	7	7	0,61	0,65	0,53	0,62	0,60	0,69	0,49	0,64
U	3	7	10	0,65	0,70	0,56	0,68	0,64	0,75	0,52	0,70
NU	1	4	4	0,51	0,55	0,49		0,51	0,55	0,48	
NU	1	4	7	0,57	0,62	0,50		0,56	0,61	0,48	
NU	1	4	10	0,61	0,65	0,51		0,60	0,66	0,49	
NU	1	7	4	0,50	0,55	0,49		0,50	0,55	0,48	
NU	1	7	7	0,54	0,61	0,50		0,54	0,61	0,48	
NU	1	7	10	0,57	0,64	0,50		0,56	0,65	0,49	
NU	2	4	4	0,50	0,54	0,48	0,53	0,49	0,53	0,46	0,51
NU	2	4	7	0,55	0,63	0,50	0,59	0,53	0,62	0,47	0,57
NU	2	4	10	0,59	0,67	0,52	0,63	0,57	0,67	0,49	0,61
NU	2	7	4	0,51	0,57	0,48	0,54	0,50	0,57	0,46	0,53
NU	2	7	7	0,56	0,65	0,49	0,59	0,55	0,66	0,46	0,58
NU	2	7	10	0,61	0,70	0,51	0,63	0,59	0,72	0,48	0,63
NU	3	7	4	0,50	0,57	0,46	0,56	0,49	0,57	0,44	0,54
NU	3	7	7	0,56	0,66	0,48	0,64	0,54	0,66	0,45	0,63
NU	3	7	10	0,61	0,72	0,51	0,69	0,59	0,73	0,47	0,69

Notes
U : Uniforme, NU : Non uniforme
Pour l'ensemble des scénarios de ce tableau, les écart-types des 500 AUROC oscillaient entre 0,04 et 0,05.

Suite à nos résultats avec les erreurs de Guttman, on s'est demandé si d'autres indicateurs de *fit* (notamment des indicateurs paramétriques) pourraient être plus fins que les erreurs de Guttman et présenter de meilleures performances. Nous avons donc cherché à développer une méthode de détection "alternative", qui serait cette fois-ci basée sur des indicateurs de *fit* paramétriques : les indices INFIT et OUTFIT.

4.5 Alternative avec les indices INFIT et OUTFIT

Jusqu'ici, nous avons utilisé les erreurs de Guttman pour mesurer les incohérences dans les réponses des patients. Cependant, d'autres mesures paramétriques de *person fit* existent pour mesurer la cohérence des réponses des individus. On peut en particulier citer les indices INFIT (*inlier-sensitive fit*) et OUTFIT (*outlier-sensitive fit*).

Ces indices peuvent être calculés pour chaque individu à la suite de l'estimation d'un modèle de l'IRT ou de la RMT. On les obtient à partir des résidus du modèle. Les résidus ainsi que les indices INFIT et OUTFIT sont définis ci-dessous dans le cadre d'un PCM estimé à partir de J items ayant M $(0, ..., M-1)$ modalités de réponse.

Pour chaque individu i et chaque item j, on définit le résidu r_{ij} comme l'écart entre la réponse de l'individu (notée X_{ij}) et la prédiction de cette réponse par le modèle (notée $E(X_{ij})$) :

$$\begin{aligned} r_{ij} &= X_{ij} - E(X_{ij}) \\ &= X_{ij} - \sum_{p=0}^{M-1} p \times P(X_{ij} = p \mid \widehat{\theta_i}, \widehat{\delta}_{j1}, ..., \widehat{\delta}_{jM-1}) \end{aligned} \tag{4.4}$$

$\widehat{\theta_i}$ désigne la prédiction du niveau de la variable latente de l'individu i et les $\widehat{\delta}_{j1}, ..., \widehat{\delta}_{jM-1}$ sont les estimations des paramètres de l'item j.

Les indices INFIT et OUTFIT pour chaque individu i s'obtiennent ensuite grâce aux formules suivantes :

$$INFIT_i = \frac{\sum_{j=1}^{J} {r_{ij}}^2}{\sum_{j=1}^{J} Var(X_{ij})} \tag{4.5}$$

$$OUTFIT_i = \frac{1}{J} \times \sum_{j=1}^{J} \frac{{r_{ij}}^2}{Var(X_{ij})} \tag{4.6}$$

Où :

$$Var(X_{ij}) = \sum_{p=0}^{M-1} \left[(p - E(X_{ij}))^2 \times P(X_{ij} = p \mid \widehat{\theta_i}, \widehat{\delta}_{j1}, ..., \widehat{\delta}_{jM-1}) \right] \tag{4.7}$$

Ces indices sont toujours positifs. Une valeur supérieure à 1 indique un *underfit* (les réponses de l'individu sont plus aléatoires que prévu par le modèle) et une valeur inférieure à 1 indique un *overfit* (les réponses de l'individu sont moins aléatoires que prévu par le modèle) [164]. Ces deux indices ne sont pas tout à fait sensibles aux mêmes types d'incohérences. L'indice INFIT est sensible aux incohérences qui surviennent à proximité du niveau de la variable latente de l'individu. Par exemple, dans le cas d'items binaires, un individu qui donne des réponses inattendues aux items situés à proximité de son niveau de variable latente aura un indice INFIT élevé. L'indice OUTFIT est quant à lui sensible aux incohérences qui surviennent loin du niveau de la variable latente de l'individu. En reprenant le cas des items binaires, l'indice OUTFIT sera élevé pour un individu ayant un niveau de variable latente élevé, mais qui ne réussirait pas les items faciles. Idem pour un individu avec un faible niveau de variable latente qui réussirait des items difficiles.

La méthode alternative que nous avons étudiée est inspirée de celle basée sur les erreurs de Guttman, mais elle se base sur les indices INFIT et OUTFIT. Ses étapes sont décrites ci-après.

Étape n°1 : Estimation d'un PCM transversal au temps t_1

La première étape de la méthode consiste à estimer, par maximum de vraisemblance, un PCM transversal sur les données récoltées au premier temps de mesure t_1. Pour rappel, ce modèle s'écrit :

$$P\left(X_{ij}^{(t_1)} = x \mid \theta_i^{(t_1)}, \delta_{j1}^{(t_1)}, \ldots, \delta_{jM-1}^{(t_1)}\right) = \frac{\exp\left(x\theta_i^{(t_1)} - \sum_{p=1}^{x}\delta_{jp}^{(t_1)}\right)}{\sum_{l=0}^{M-1}\exp\left(l\theta_i^{(t_1)} - \sum_{p=1}^{l}\delta_{jp}^{(t_1)}\right)} \tag{4.8}$$

Afin que le modèle soit identifiable, la moyenne de la variable latente $\Theta^{(t_1)}$ a été fixée à 0 (contrainte d'identifiabilité). Suite à l'estimation de ce modèle, les estimations ponctuelles des paramètres de seuil des items $\hat{\delta}_{jp}^{(t_1)}$ sont récupérées et les indices INFIT et OUTFIT sont calculés pour chaque individu. Ils seront notés $INFIT^{(t_1)}$ et $OUTFIT^{(t_1)}$.

Étape n°2 : Estimation d'un PCM transversal au temps t_2

La deuxième étape de la méthode consiste à estimer un nouveau PCM transversal sur les données récoltées au second temps de mesure t_2. Pour ce modèle, les paramètres de seuil des items sont fixés à ceux estimés précédemment (contrainte d'invariance longitudinale) :

$$P\left(X_{ij}^{(t_2)} = x \mid \theta_i^{(t_2)}, \delta_{j1}^{(t_2)}, \ldots, \delta_{jM-1}^{(t_2)}\right) = \frac{\exp\left(x\theta_i^{(t_2)} - \sum_{p=1}^{x}\delta_{jp}^{(t_2)}\right)}{\sum_{l=0}^{M-1}\exp\left(l\theta_i^{(t_2)} - \sum_{p=1}^{l}\delta_{jp}^{(t_2)}\right)} \tag{4.9}$$

Contrainte d'invariance longitudinale : $\forall\ p$ et $\forall\ j$: $\delta_{jp}^{(t_2)} = \hat{\delta}_{jp}^{(t_1)}$

Suite à l'estimation de ce modèle, les indices INFIT et OUTFIT sont de nouveau calculés pour chaque individu. Ils sont notés $INFIT^{(t_2)}$ et $OUTFIT^{(t_2)}$.

Étape n°3 : Calculer l'évolution des indices INFIT et OUTFIT

Pour cette dernière étape, deux indicateurs sont calculés : I_{INFIT} et I_{OUTFIT}. Ces indicateurs correspondent aux évolutions respectives des indices INFIT et OUTFIT entre t_1 et t_2.

$$I_{INFIT} = INFIT^{(t_2)} - INFIT^{(t_1)} \tag{4.10}$$

$$I_{OUTFIT} = OUTFIT^{(t_2)} - OUTFIT^{(t_1)} \tag{4.11}$$

Hypothèses : Comme pour les indicateurs basés sur les erreurs de Guttman, on s'attendait à ce que les individus n'expérimentant pas de recalibration aient des indices INFIT et OUTFIT stables dans le temps : $I_{INFIT} \approx 0$ et $I_{OUTFIT} \approx 0$. Au contraire, on s'attendait à une augmentation des indices INFIT et OUTFIT pour les individus expérimentant de la recalibration (la recalibration entraînant potentiellement une augmentation des résidus entre les deux temps de mesure).

Cette méthode a été éprouvée par simulations, sur une partie des scénarios de l'étude n°2 présentée en section 4.4. Les scénarios sélectionnés étaient ceux avec $N = 200$ individus, $\Delta = 0$ (pas d'évolution de la variable latente en moyenne), $J = 4$ ou 7 items avec $M = 4$, 7 ou 10 modalités de réponse, de la recalibration uniforme simulée pour $p = 25\%$ des individus sur $J_{RS} = 1$, 2 ou 3 items en positon "Basse", "Moyenne" ou "Haute".

Comme pour les indicateurs basés sur les erreurs de Guttman (I et I_{norm}) nous avons estimé, pour chaque scénario, les AUROC moyennes associées à ces nouveaux indicateurs. Les résultats sont donnés dans le tableau 4.8 en parallèle des résultats obtenus avec les indicateurs basés sur les erreurs de Guttman I et I_{norm}.

TABLEAU 4.8 – Moyennes des 500 aires sous la courbe ROC (AUROC) associées aux indicateurs I_{INFIT}, I_{OUTFIT}, I et I_{norm} en fonction du nombre d'items touchés par la recalibration (J_{RS}), du nombre total d'items (J), du nombre de modalités de réponse (M) et de la position des items touchés par la recalibration (Moyenne, Basse ou Haute).
Les scénarios considérés dans ce tableau sont ceux où la variable latente est en moyenne stable entre t_1 et t_2 (c.-à-d. $\Delta = 0$)

Δ = 0				**AUROC moyennes associées à I_{INFIT}**			**AUROC moyennes associées à I_{OUTFIT}**			**AUROC moyennes associées à I**			**AUROC moyennes associées à I_{norm}**		
				Position :			Position :			Position :			Position :		
Forme RC	J_{RS}	J	M	Moy.	Basse	Haute	Moy.	Basse	Haute	Moy.	Basse	Haute	Moy.	Basse	Haute
U	1	4	4	0,54	0,55	0,51	0,54	0,56	0,50	0,53	0,56	0,50	0,53	0,56	0,49
U	1	4	7	0,58	0,59	0,52	0,58	0,61	0,51	0,58	0,60	0,52	0,57	0,62	0,50
U	1	4	10	0,62	0,63	0,54	0,62	0,65	0,52	0,62	0,63	0,54	0,61	0,66	0,51
U	1	7	4	0,52	0,54	0,50	0,52	0,55	0,49	0,52	0,54	0,50	0,52	0,55	0,49
U	1	7	7	0,56	0,58	0,51	0,55	0,60	0,49	0,56	0,58	0,50	0,55	0,60	0,49
U	1	7	10	0,59	0,62	0,53	0,58	0,64	0,50	0,59	0,61	0,52	0,58	0,64	0,51
U	2	4	4	0,53	0,55	0,50	0,53	0,57	0,49	0,53	0,55	0,50	0,53	0,57	0,49
U	2	4	7	0,57	0,61	0,54	0,57	0,63	0,51	0,56	0,61	0,52	0,56	0,64	0,50
U	2	4	10	0,62	0,64	0,56	0,62	0,68	0,54	0,61	0,65	0,56	0,60	0,69	0,52
U	2	7	4	0,54	0,57	0,50	0,54	0,59	0,48	0,54	0,57	0,49	0,53	0,59	0,48
U	2	7	7	0,59	0,63	0,52	0,58	0,66	0,50	0,58	0,63	0,52	0,58	0,67	0,50
U	2	7	10	0,64	0,68	0,54	0,63	0,71	0,51	0,63	0,67	0,53	0,62	0,72	0,51
U	3	7	4	0,54	0,58	0,50	0,54	0,60	0,48	0,54	0,58	0,49	0,54	0,61	0,47
U	3	7	7	0,60	0,64	0,53	0,59	0,68	0,50	0,59	0,64	0,52	0,59	0,69	0,49
U	3	7	10	0,64	0,69	0,56	0,63	0,74	0,52	0,63	0,69	0,55	0,63	0,75	0,52

Notes :

RC : Recalibration, U : Uniforme

Pour l'ensemble des scénarios de ce tableau, les écart-types des 500 AUROC oscillaient entre 0,04 et 0,05.

Pour les scénarios où $M \geq 7$, les estimations des paramètres de seuil des items étaient quasiment systématiquement désordonnées.

Les AUROC moyennes de ces deux nouveaux indicateurs I_{INFIT} et I_{OUTFIT} se sont révélées très proches de celles des indicateurs I et I_{norm}. En termes de capacité à discriminer les individus avec et sans recalibration simulée, ces deux nouveaux indicateurs ne permettent donc pas de faire mieux que les indicateurs basés sur les erreurs de Guttman.

4.6 Bilan

Les indices de *person fit*, comme le nombre d'erreurs de Guttman ou les indices INFIT et OUTFIT, semblaient être une piste intéressante pour la détection de la recalibration à un niveau individuel. Néanmoins, les résultats obtenus par simulations indiquent qu'ils ne permettent pas de distinguer les individus pour lesquels on a simulé de la recalibration des autres individus (pour lesquels on n'en a pas simulé). Il existe bien sûr d'autres indices de *person fit*, mais il est probable qu'ils ne présentent pas de meilleures performances que ceux étudiés dans ce manuscrit.

Les travaux présentés dans ce chapitre ont soulevé de nombreuses questions méthodologiques et conceptuelles. Comme mentionné précédemment, on s'est tout d'abord questionné sur la façon de représenter les résultats d'une étude de simulation où l'on s'intéresse à une granularité individuelle. Cette représentation a été un véritable écueil : notre choix initial d'utiliser la moyenne des indicateurs nous ayant fait perdre cette granularité individuelle, mais également induit en erreur sur les performances de nos indicateurs. En effet, si la recalibration s'accompagne d'une augmentation du nombre d'erreurs de Guttman (ou des indices INFIT/OUTFIT), la capacité discriminante des indicateurs étudiés restait faible (voir parfois nulle). C'est pourquoi nous nous sommes ensuite focalisés sur les aires sous la courbe ROC. Les AUROC nous ont semblées être un bon compromis, puisqu'elles permettaient de quantifier la capacité discriminante des indicateurs, en évaluant le chevauchement de leurs distributions dans les deux groupes d'individus simulés (avec ou sans recalibration).

Nous nous sommes également questionnés sur la façon de simuler de la recalibration à un niveau individuel. Dans nos deux études de simulation, nous avons opérationnalisé la recalibration en décalant les paramètres de seuil de certains items entre les deux temps de mesure. Les

paramètres décalés étaient les mêmes pour tous les individus pour qui on simulait de la recalibration, tout comme la taille et la direction de ces décalages. Nous avons donc supposé que tous les individus expérimentaient tous exactement la même recalibration. Conceptuellement, on peut néanmoins se demander si cette façon de simuler était adéquate, et dans le cas contraire, quelle est la bonne façon de simuler de la recalibration à un niveau individuel. Nous n'avons, pour l'heure, pas la réponse à cette question.

Une autre piste pour la détection du *response shift* au niveau des items et avec une granularité individuelle a été récemment suggérée par Sawatzky [157] : l'utilisation des *Latent variable mixture models* (LVMM). Ce type de modèle combine un modèle à classes latentes avec un modèle à variable latente (modèles de la famille de Rasch ou de Lord). L'idée générale de cette méthode est d'identifier différents groupes latents (classes latentes) qui expérimentent le *response shift* différentiellement. Pour ce faire, un modèle à variable latente longitudinal est estimé entre deux temps de mesure. Au sein de ce modèle, différentes classes latentes sont spécifiées :

- Une première classe latente regroupant des individus n'expérimentant pas de *response shift*. Dans cette classe, les paramètres des items sont contraints à être invariants, c'est-à-dire constants entre les deux temps de mesure.
- Une (ou plusieurs) classe(s) latente(s) regroupant des individus expérimentant du *response shift*. Dans ces classes, certains paramètres d'item sont autorisés à varier au cours du temps. On pourrait par exemple imaginer une classe où les individus expérimentent de la recalibration sur l'item j et une autre classe où les individus expérimentent de la recalibration sur l'item j' (ou alors deux classes où les individus expérimentent de la recalibration se manifestant sur le même item, mais dans des directions différentes dans chacune des classes).

Les probabilités d'appartenance à chacune des classes pourraient ensuite être utilisées pour quantifier la susceptibilité de chaque individu à expérimenter du *response shift*. Les caractéristiques des individus associées aux classes pourraient également être utilisées pour décrire les individus susceptibles d'expérimenter du *response shift*.

Cette méthode pourrait être prometteuse pour la recherche sur le *response shift*, car elle

permet de s'intéresser à l'hétérogénéité de ce phénomène. Une application sur données réelles a été présentée au congrès ISOQOL 2021 [165] par Sawatzky *et al.*, mais cette méthode n'a pour l'heure fait l'objet d'aucune publication (à notre connaissance). Des recherches supplémentaires, notamment avec des études de simulation, sont nécessaires pour évaluer les performances d'une telle méthode. De plus, de nombreuses questions se posent :

- Comment déterminer le nombre de classes adéquat ?
- Quels paramètres d'item libérer dans chacune des classes avec du *response shift* ? Sur quel critère ce choix se base-t-il ?

Les LVMM ont déjà été utilisés à plusieurs reprises pour détecter du DIF lors de l'analyse de données auto-rapportées (en santé et dans les sciences de l'éducation) [166–174]. Conceptuellement, cette méthode semble intéressante, car elle permettrait de détecter du DIF sans faire d'hypothèses *a priori* sur les covariables à l'origine du fonctionnement différentiel de certains items. En pratique, les LVMM cherchent à former des classes latentes de sorte que :

(i) Les individus d'une même classe partagent la même perception du questionnaire ;

(ii) Un ou plusieurs items fonctionne(nt) différemment entre deux classes latentes.

En 2011, une étude de simulation réalisée par DeMars et Lau a questionné le recours aux LVMM pour la détection du DIF [175]. Ces auteures ont simulé deux groupes de 3 000 individus chacun répondant à un questionnaire composé de J = 10 ou 20 items binaires dont 4 (ou 8) fonctionnaient différemment entre les deux groupes. Elles ont ensuite cherché à déterminer si un LVMM avec deux classes latentes (entre lesquelles les paramètres d'item pouvaient varier) réussissait à retrouver les classes simulées. Leurs résultats ont indiqué que les classes latentes qui émergeaient n'étaient pas fortement associées aux véritables classes sous-jacentes simulées. Une autre étude de simulation a été récemment réalisée par Sajobi *et al.* [176]. Parmi les scénarios explorés, ces auteurs ont simulé deux groupes de tailles déséquilibrées (N_1 = 225 et N_2 = 675 individus) répondant à un questionnaire composé de $J = 12$ ou 30 items polytomiques (dont 25 à 50% fonctionnaient différemment entre ces deux groupes). Ils ont ensuite cherché à déterminer si

un LVMM à deux classes latentes réussissait à retrouver les classes simulées. Idem avec un modèle avec 3 classes latentes. Si les auteurs indiquaient que les taux de recouvrement des classes étaient assez élevés, ils rappelaient néanmoins que ce type de modèle n'était pas forcément robuste à une mauvaise spécification du nombre de classes latentes (dans leur cas : chercher trois classes au lieu de deux).

D'un point de vue computationnel, il y a un nombre très important de paramètres à estimer dans les LVMM. Ils nécessitent donc de grands échantillons. De plus, ces modèles peuvent présenter des problèmes de convergence : non-convergence, convergence sur un *extremum* local, ou encore obtention d'estimations aberrantes. Enfin, on peut également se questionner sur leurs performances en cas de classes déséquilibrées (comme c'est probablement le cas lors de la recherche de *response shift*).

Face à ces observations, il semble capital de déterminer, par une étude de simulation, si un modèle LVMM parfaitement spécifié (bon nombre de classes latentes et bon paramètres d'item libérés) permet de retrouver les classes simulées. Nous avons effectué des premiers tests en ce sens avec le logiciel Mplus, mais de nombreuses difficultés computationnelles ont été rencontrées. Le temps d'estimation d'un PCM longitudinal avec deux classes latentes est de l'ordre de plusieurs heures.

Suite à l'ensemble de ces observations, nous avons décidé de ne pas poursuivre pour ces travaux (sur les indices de *fit* et sur les modèles LVMM). Néanmoins, comme évoqué dans l'introduction, une autre piste pourrait permettre d'étudier la variabilité interindividuelle du *response shift* : l'introduction de covariables dans la procédure de détection du *response shift*. Ce type de méthode nécessite des développements méthodologiques. En effet, il s'agirait ici d'étudier les changements dans l'interprétation des items d'un questionnaire au cours du temps, en considérant plusieurs covariables. Pour pouvoir mener ce type d'analyse, il faut déjà être capable de prendre correctement en compte les différences d'interprétation des items qui pourraient être induites par ces covariables au premier temps de mesure. La suite de ma thèse s'est donc en

partie focalisée sur cette problématique, en cherchant à détecter du DIF en présence de deux covariables (chapitre 5).

Par ailleurs, lorsque le *response shift* fait suite à un événement défiant hautement les capacités des individus, il se pourrait qu'il soit en partie expliqué par la survenue de développement post-traumatique [18]. Aussi, il pourrait être intéressant de s'intéresser aux liens entre ces deux phénomènes et déterminer, par exemple, si l'on retrouve du *response shift* chez des individus ayant expérimenté du développement post-traumatique. Ce type d'études permettrait d'approcher une granularité de détection du *response shift* assez fine. Pour pouvoir mener à bien ce type de recherche, il est nécessaire d'avoir un outil de mesure fiable et valide du développement post-traumatique. Le dernier travail de cette thèse porte donc sur l'évaluation des propriétés psychométriques d'une version française de l'inventaire du développement post-traumatique (chapitre 6).

Chapitre 5

Détection du DIF avec deux covariables binaires : extension de ROSALI

Sommaire

5.1 Motivations et objectifs

Comme on l'a vu dans le chapitre 3, le fonctionnement différentiel des items (*Differential Item Funcitoning*, DIF) est un phénomène important, qui doit être pris en compte lors de l'analyse des données rapportées par les patients. En effet, ignorer le DIF peut nuire à l'interprétation des données rapportées par les patients et à la décision médicale qui en découle. De plus, le DIF est également un phénomène d'intérêt, puisqu'il peut être intéressant de chercher à comprendre pourquoi certains items fonctionnent différemment chez certains individus.

Lorsque l'on analyse des données rapportées par les patients, les variables à l'origine d'un fonctionnement différentiel d'items peuvent être multiples [21]. La perception des items peut par exemple différer en fonction du genre du répondant, de son âge, mais également de ses caractéristiques cliniques [177]. Il serait donc intéressant de pouvoir prendre en compte différentes covariables lors de la détection du DIF.

La solution la plus directe serait de rechercher du DIF pour chacune des covariables d'intérêt, une covariable à la fois. L'analyse aurait donc lieu de façon séparée et indépendante. Cette solution ne vient cependant pas sans d'importants inconvénients. En effet, considérons le cas où deux covariables corrélées (notées C_1 et C_2) sont suspectées d'induire du DIF, mais qu'une seule de ces deux covariables, la covariable C_1, en induit réellement. Mener la détection du DIF pour chacune des covariables de façon séparée et indépendante risque ici de mener à une conclusion erronée. On pourrait par exemple incriminer à tort la covariable C_2 en plus de la covariable C_1. Dans notre exemple, cette conclusion erronée serait la conséquence de la corrélation entre les deux variables : on attribue à la covariable C_2 le fonctionnement différentiel causé par la covariable C_1. Ainsi, pour pouvoir démêler les sources du DIF, la mise en œuvre séparée et indépendante des méthodes à une covariable ne semble pas être appropriée [19].

Une méthode de détection du DIF permettant de prendre simultanément en compte plusieurs covariables, éventuellement corrélées, pourrait être d'un grand intérêt pour mieux identifier les variables à l'origine du fonctionnement différentiel des items. Plusieurs approches statistiques ba-

sées sur la théorie de la réponse aux items ou la théorie de la mesure de Rasch ont été récemment développées dans ce but :

- La procédure IRT-C proposée par Tay *et al.* [20, 80, 84] ;
- La procédure MIMIC de Chun *et al.* [22] ;
- La procédure DIF-IFT de Bollman *et al.* [71] ;
- La procédure (G)PCM-Lasso de Schauberger et Mair [19].

Ces méthodes ont été présentées en détail dans le chapitre 3. Leur principe général et les limites qu'elles présentent sont rappelées dans la figure 5.1.

Dans le même temps, l'algorithme ROSALI avec une covariable binaire [150] a été développé pour détecter la non-invariance de la mesure entre groupes et au cours du temps (voir chapitre 3). Dans sa version actuelle, ROSALI permet de considérer deux temps de mesure et vise à estimer l'effet d'une covariable binaire ("effet groupe") sur le fonctionnement des items, le *response shift* (recalibration) et le changement au cours du temps de la variable latente. L'algorithme est composé de deux parties. La première est dédiée à la détection de la non-invariance de la mesure au premier temps de mesure entre les deux groupes d'individus définis par la covariable binaire (assimilable à une détection du DIF). La seconde partie de l'algorithme est quant à elle dédiée à la détection de la non-invariance de la mesure au cours du temps (c.-à-d. détection du *response shift*). Un modèle final permet d'estimer l'effet de la covariable sur :

(i) le niveau moyen de la variable latente au premier temps de mesure ;

(ii) l'évolution moyenne de la variable latente entre les deux temps de mesure.

Ces estimations tiennent compte du DIF et de la recalibration mis en évidence au cours de l'algorithme. Des simulations ont démontré que ROSALI n'inférait pas du DIF à tort lorsque aucun fonctionnement différentiel n'a été simulé [178]. Cependant, ses performances pour détecter du DIF n'ont jamais été évaluées. À ce jour, il y a une volonté d'étendre l'algorithme ROSALI pour considérer simultanément plusieurs sources de non-invariance (par exemple : le sexe, le pays, ou encore une caractéristique clinique).

Méthode	Principe général	Limites
IRT-C *Tay et al.*	• On suspecte une covariable d'induire du DIF sur un item si le résidu bivarié associé à cette paire item-covariable est élevé. • Suspicion confirmée ou infirmée par un test statistique (Wald) • Méthode itérative : modèle mis à jour quand du DIF est détecté	• Pas d'extension aux items polytomiques • Utilisation des résidus bivariés remise en question
MIMIC *Chun et al.*	• Détection du DIF basée sur un test du rapport de vraisemblance comparant : (i) Un modèle sans DIF sur l'item (ii) Un modèle où toutes les covariables incluses dans l'analyse induisent du DIF sur l'item • Items testés un par un • Pas de mise à jour itérative du modèle	• Une étude de simulation a montré que cette méthode rencontrait des difficultés à identifier les covariables à l'origine du DIF • Suppose que l'effet des covariables sur les paramètres de seuil est homogène (DIF homogène)
DIF-IFT *Bollmann et al.*	Recherche hiérarchique du DIF basée sur des tests de permutation	• Ne semble pas permettre de modéliser l'effet des covariables sur la variable latente • Absence de données sur les performances pour détecter du DIF : - Quand plusieurs covariables en sont à l'origine - Quand les covariables introduites dans l'analyse sont corrélées
(G)PCM-Lasso pour DIF H *Schauberger et al.* (G)PCM-Lasso pour DIF H et NH *Schauberger et al.*	• La détection du DIF est ramenée à un problème de sélection de paramètres à estimer. • Cette sélection est réalisée grâce à une estimation pénalisée (ou régularisée) du modèle de mesure	• Suppose que le DIF est homogène • N'a jamais été évaluée par simulations

FIGURE 5.1 – Méthodes de détection du DIF permettant de considérer simultanément plusieurs covariables

Notes :

IRT-C : Procédure "Item Response Theory with Covariates"

MIMIC : Procédure "Multiple Indicators Multiple Causes"

DIF-IFT : Procédure "Item-Focused Trees to detect DIF based on Partial Credit Models"

(G)PCM-Lasso : Procédure "Lasso regularization approach for the detection of differential item functioning in (Generalized) Partial Credit Models"

DIF H : Fonctionnement différentiel des items homogène, DIF NH : Fonctionnement différentiel des items non homogène

Les travaux présentés dans ce chapitre visent à proposer une extension de la partie 1 de l'algorithme ROSALI, pour pouvoir prendre en compte deux covariables binaires. L'objectif de cette extension est d'être en capacité de détecter les items affectés par du DIF et les covariables qui en sont à l'origine, afin d'ajuster les comparaisons des groupes en conséquence. Les performances de cette extension ont été évaluées à l'aide d'une étude de simulation. Ces performances ont été comparées à celles de la méthode proposée par Schauberger et Mair qui permet la détection du DIF homogène ou non homogène grâce à l'estimation pénalisée d'un PCM (pénalisation de type *lasso*) [19]. Cette méthode a été sélectionnée, car elle repose sur une philosophie très différente de celle de ROSALI (contrairement aux autres méthodes identifiées dans la littérature, qui basent la détection du DIF sur la réalisation de tests statistiques).

Cette étude de simulation a fait l'objet d'un article qui est actuellement en cours de resoumission. La version actuelle de cet article est disponible en annexe (cf. annexe A).

5.2 Partie 1 de l'algorithme ROSALI (avec une covariable binaire)

La partie 1 de ROSALI est un algorithme itératif, qui vise à détecter de la non-invariance de la mesure entre deux groupes d'individus à un temps de mesure donné. Ces deux groupes sont définis par la covariable binaire introduite dans l'analyse. Cet algorithme s'appuie sur l'estimation successive de différents PCM transversaux, où un effet de la covariable est modélisé à la fois sur le niveau moyen de la variable latente et sur les paramètres de seuil de certains items. Ainsi, la partie 1 de ROSALI peut être vue comme une approche MIMIC. La formulation générique des PCM estimés au cours de cet algorithme est donnée ci-dessous :

$$P(X_{ij} = x \mid \theta_i, \delta_{j1}, .., \delta_{jM_j-1}, C_i, \beta, \gamma_{j1}, .., \gamma_{jM_j-1}) = \frac{\exp\left(x\left[\theta_i + \beta C_i\right] - \sum_{p=1}^{x}\left[\delta_{jp} + \gamma_{jp} C_i\right]\right)}{\sum_{l=0}^{M_j-1}\exp\left(l\left[\theta_i + \beta C_i\right] - \sum_{p=1}^{l}\left[\delta_{jp} + \gamma_{jp} C_i\right]\right)} \tag{5.1}$$

Où :

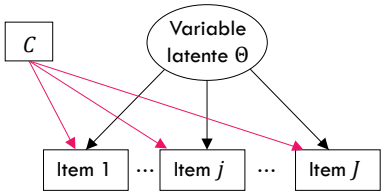
- X_{ij} est la variable représentant la réponse de l'individu i à l'item j ($j = 1, \ldots, J$). Les valeurs possibles pour X_{ij} sont tous les entiers compris entre 0 et $M_j - 1$ où M_j désigne le nombre de modalités de l'item j.
- θ_i désigne la valeur de la variable latente Θ pour l'individu i. Pour rappel, cette variable Θ est supposée distribuée selon une loi normale.
- C_i correspond à la valeur de la variable binaire C pour l'individu i. Les valeurs possibles sont 0 ou 1.
- Les δ_{jp} désignent les paramètres de seuil de l'item j. Il y a un paramètre pour chaque modalité de réponse p supérieure à 0 ($p = 1, \ldots M_j - 1$).
- β est l'effet de la covariable binaire C sur le niveau moyen de la variable latente. Ce coefficient correspond à la différence entre μ_1 et μ_0, les moyennes respectives de la variable latente Θ dans les groupes $C = 1$ et $C = 0$.
- Les γ_{jp} désignent les paramètres associés au DIF induit par la covariable C. Ces coefficients opérationnalisent l'écart entre les paramètres de seuil de l'item j dans le groupe $C = 0$ et les paramètres de seuil de l'item j dans le groupe $C = 1$:
  - Dans le groupe $C = 0$, les paramètres de seuil de l'item j correspondent aux δ_{jp} ;
  - Dans le groupe $C = 1$, les paramètres de seuil de l'item j sont égaux à $\delta_{jp} + \gamma_{jp}$.

  Si la covariable C n'induit pas de DIF sur l'item j (ou si on ne modélise pas cet effet), alors les paramètres γ_{jp} sont nuls.

Comme la partie 1 de ROSALI est basée sur l'estimation de PCM, elle permet uniquement de détecter des différences entre groupes au niveau des paramètres de seuil des items. Les étapes de cet algorithme sont décrites ci-dessous (voir le chapitre 3, figure 3.15 pour une représentation graphique).

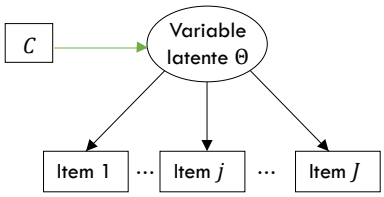
Étape 1 : Estimation d'un modèle complétement non-invariant (Modèle A)

Cette première étape vise à estimer un PCM transversal sans contraintes d'invariance. Dans ce modèle, la covariable C induit du DIF sur tous les items. Ainsi, tous les paramètres de DIF γ_{jp} sont librement estimés dans l'équation (5.1) (représentés ci-dessous par les flèches roses). Pour que le modèle soit identifiable, l'effet de la covariable C sur le niveau de la variable latente est contraint à être nul ($\beta = 0$).



Étape 2 : Estimation d'un modèle complétement invariant (Modèle B)

Cette seconde étape vise à estimer un PCM transversal avec des contraintes d'invariance sur tous les items. Dans ce modèle, il n'y a pas de DIF modélisé. Ainsi, tous les paramètres de DIF γ_{jp} sont contraints à être nuls dans l'équation (5.1) ($\forall j$ et p). L'effet de la covariable C sur le niveau de la variable latente peut cette fois-ci être estimé.



Étape 3 : Test de la présence globale de DIF

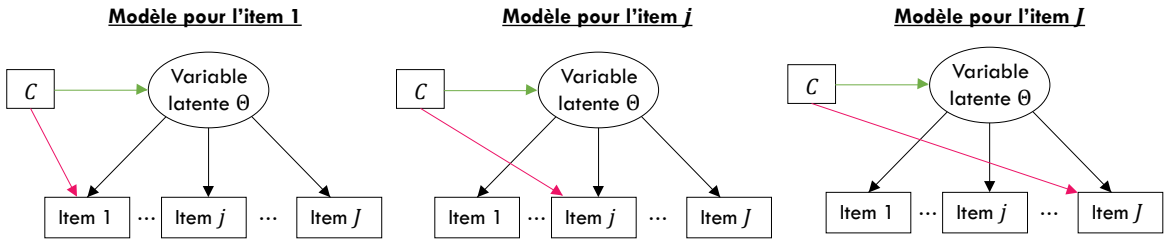
Cette troisième étape vise à évaluer la présence globale de DIF en comparant les modèles A et B à l'aide d'un test du rapport de vraisemblance (*likelihood-ratio test*, LRT) avec un risque de première espèce fixé à $\alpha = 5\%$. Ce test est possible puisque le modèle B est emboîté dans le modèle A. Si ce test n'est pas significatif, alors on suppose que la covariable C n'induit pas de DIF, l'algorithme passe alors directement à l'étape 5 où un modèle final est estimé (ce modèle final

correspond au modèle B sans DIF). Dans le cas contraire, on suppose que certains paramètres d'items sont différents entre les groupes et l'algorithme passe à l'étape suivante (l'étape 4).

Étape 4 : Détection des items affectés par du DIF et identification de la forme du DIF

Cette étape est itérative et a pour objectif d'identifier les items affectés par du DIF ainsi que la forme de DIF sous-jacente. Tous les items sont à l'étude pour la détection du DIF. Pour mener à bien cette détection, un nouveau modèle (modèle C) est introduit de façon à ce que le modèle C corresponde au modèle B au début de l'étape.

À partir du modèle C, l'algorithme estime de nouveaux PCM (autant de modèles que d'items à tester). Pour chacun de ces modèles, la contrainte d'invariance associée à l'item testé est relâchée et les autres contraintes d'invariance restent inchangées.



À partir de chacun de ces nouveaux modèles, un test statistique pour hypothèses jointes est réalisé pour déterminer si la covariable C induit ou non un fonctionnement différentiel significatif pour l'item étudié. Les hypothèses nulle et alternative de ce test sont :

$$\begin{aligned} &\mathrm{H}_0 : \forall\ p, \gamma_{jp} = 0 \quad \text{(La covariable } C \text{ n'induit pas de DIF sur l'item } j) \\ &\mathrm{H}_1 : \exists\ p : \gamma_{jp} \neq 0 \quad \text{(La covariable } C \text{ induit du DIF sur l'item } j) \end{aligned} \tag{5.2}$$

Le modèle associé au test le plus significatif (plus petite *p-value*) après une correction de Bonferroni est retenu. Le risque de première espèce est fixé à $\alpha = 5\%$/nombre d'items. On supposera que l'item associé à ce modèle, celui pour qui la contrainte d'invariance a été relâchée, est affecté

par du DIF. Cet item est noté item j^* pour la suite de l'algorithme. Si aucun test n'est significatif, l'algorithme passe alors à l'étape 5. Sinon, à partir du modèle retenu, la forme du DIF est déterminée par un nouveau test statistique pour hypothèses jointes, réalisé avec un risque de première espèce à 5%. Les hypothèses nulle et alternative de ce test sont :

$$\begin{aligned} &\mathrm{H}_0 : \forall\ p,\ \gamma_{j^*p} = \gamma_{j^*} \text{ (DIF homogène)} \\ &\mathrm{H}_1 : \exists\ p,\ p' :\ \gamma_{j^*p} \neq \gamma_{j^*p'} \text{ (DIF non homogène)} \end{aligned} \tag{5.3}$$

Le modèle C est ensuite mis à jour pour prendre en compte le DIF qui vient d'être mis en évidence. L'item ne sera plus testé.
Si le test précédent est significatif, on suppose alors que le DIF est non homogène et les paramètres de DIF γ_{j^*p} associés à la paire retenue sont librement estimés.
Si le test est non significatif, on considère que le DIF est homogène et les paramètres de DIF γ_{j^*p} sont estimés, mais contraints à être constants : $\forall\ p,\ \gamma_{j^*p} = \gamma_{j^*}$.

L'étape 4 est répétée sur les items restants jusqu'à ce qu'il n'y ait plus d'item retenu (c.-à-d., plus de DIF mis en évidence) où jusqu'à ce qu'il ne reste plus qu'un seul item à tester.

Théoriquement, la partie 1 de ROSALI s'arrête ici, et on enchaîne sur la partie 2 de l'algorithme, dédiée à la recalibration. Comme nous sommes dans un cadre transversal, une cinquième étape a été ici rajoutée pour estimer l'effet de la covariable C sur le niveau moyen de la variable latente (ajusté sur le DIF éventuellement mis en évidence).

Étape 5 : Estimation de l'effet de la covariable C sur le niveau moyen de la variable latente (Modèle D)

Au cours de cette dernière étape, un modèle final (modèle D) est estimé. Si l'algorithme n'a pas mis évidence de DIF alors le modèle D correspond au modèle B. Sinon, le modèle D correspond à la dernière version du modèle C. Ce modèle permet d'obtenir l'estimation de l'effet de la covariable C sur le niveau moyen de la variable latente. Cet effet est ajusté sur le DIF éventuellement détecté au cours de l'étape précédente.

5.3 Extension de la partie 1 de l'algorithme ROSALI

Comme on l'a fait pour une variable binaire C dans la section précédente, deux covariables binaires C_1 et C_2, ayant à la fois un effet sur le niveau moyen de la variable latente et sur les paramètres de seuil des items, peuvent être introduites dans un PCM :

$$P\left(X_{ij}=x \mid \theta_i,\delta_{j1},\ldots,\delta_{jM_j-1},C_{1_i},C_{2_i},\beta_1,\beta_2,\gamma_{j1}^{(C_1)},\ldots,\gamma_{jM_j-1}^{(C_1)},\gamma_{j1}^{(C_2)},\ldots,\gamma_{jM_j-1}^{(C_2)}\right)= \frac{\exp\left(x\left[\theta_i+\beta_1 C_{1_i}+\beta_2 C_{2_i}\right]-\sum_{p=1}^{x}\left[\delta_{jp}+\gamma_{jp}^{(C_1)}C_{1_i}+\gamma_{jp}^{(C_2)}C_{2_i}\right]\right)}{\sum_{l=0}^{M_j-1}\exp\left(l\left[\theta_i+\beta_1 C_{1_i}+\beta_2 C_{2_i}\right]-\sum_{p=1}^{l}\left[\delta_{jp}+\gamma_{jp}^{(C_1)}C_{1_i}+\gamma_{jp}^{(C_2)}C_{2_i}\right]\right)} \tag{5.4}$$

Les notations de ce modèle sont similaires à celles du modèle décrit dans l'équation (5.1). C'est sur ce dernier modèle que l'on s'est appuyé pour étendre la partie 1 de ROSALI. Deux versions d'extension ont été proposées. Leurs étapes sont décrites ci-dessous.

5.3.1 ROSALI-DIF : 1ère version de l'extension

ROSALI-DIF : 1re version de l'extension

Étape 1. Estimation d'un modèle complètement non-invariant (Modèle A)

Cette première étape vise à estimer un PCM transversal sans contraintes d'invariance. Dans ce modèle, les covariables C_1 et C_2 induisent du DIF sur tous les items. Ainsi, pour ce modèle, tous les paramètres de DIF $\gamma_{jp}^{(C_1)}$ et $\gamma_{jp}^{(C_2)}$ de l'équation (5.4) sont librement estimés ($\forall$ j et p). Pour que le modèle soit identifiable, l'effet des covariables sur le niveau moyen de la variable latente est contraint à être nul ($\beta_1 = \beta_2 = 0$).

Étape 2. Estimation d'un modèle complètement invariant (Modèle B)

Cette deuxième étape vise à estimer un PCM transversal invariant (c.-à-d., supposant l'absence de DIF). Ainsi, sur la base de l'équation (5.4), tous les paramètres de DIF $\gamma_{jp}^{(C_1)}$ et $\gamma_{jp}^{(C_2)}$ sont contraints à être nuls ($\forall$ j et p). L'effet des covariables sur le niveau de la variable latente (β_1 et β_2) peut cette fois-ci être estimé.

Étape 3. Test de la présence globale de DIF

La présence globale de DIF est évaluée en comparant les modèles A et B grâce à un test du rapport de vraisemblance (risque de première espèce fixé à 5%). Si ce test n'est pas significatif, alors on suppose que les deux covariables n'induisent pas de DIF et l'algorithme passe directement à l'étape 6 où le modèle final correspond au modèle B. Dans le cas contraire, l'algorithme passe à l'étape 4.

Étape 4. Identifier les paires item-covariable candidates pour la détection du DIF

Cette étape est une étape de *screening* qui vise à identifier les paires item-covariable candidates pour la détection du DIF. À partir des estimations du modèle A (modèle où les deux covariables induisent du DIF sur tous les items), un test statistique est réalisé pour chaque paire (item j, covariable C) afin de déterminer si la covariable C (C_1 ou C_2) induit significativement du DIF sur l'item j. Les hypothèses nulle et alternative de ce test sont :

$$\begin{aligned} &\mathrm{H}_0 : \forall\ p, \gamma_{jp}^{(C)} = 0 \quad \text{(La covariable } C \text{ n'induit pas de DIF sur l'item } j\text{)} \\ &\mathrm{H}_1 : \exists\ p : \gamma_{jp}^{(C)} \neq 0 \quad \text{(La covariable } C \text{ induit du DIF sur l'item } j\text{)} \end{aligned} \tag{5.5}$$

Les paires candidates pour la détection du DIF sont celles associées à un test significatif (risque de première espèce à 5%). On suppose que les autres paires sont invariantes (paires dites *anchor*). Si aucune paire n'est considérée comme candidate, alors l'algorithme passe directement à l'étape 6 où le modèle final est le modèle B.

Étape 5. Sélection des paires item-covariable affectées par du DIF parmi les paires candidates et évaluation de la forme de DIF impliquée

Cette étape itérative a pour objectif de sélectionner les paires item-covariable affectées par du DIF parmi les paires candidates et déterminer la forme de DIF sous-jacente. Un nouveau modèle, appelé modèle C, est introduit de sorte que le modèle C corresponde au modèle B (modèle sans DIF) au début de l'étape 5.

– À partir du modèle C, l'algorithme estime pour chaque paire candidate (item j, covariable C) un nouveau modèle où (i) la contrainte d'invariance associée à la paire étudiée est relâchée (estimation libre des paramètres de DIF $\gamma_{jp}^{(C)}$) et (ii) les contraintes pour les autres paires restent inchangées. À partir de chacun de ces nouveaux modèles, un test statistique est réalisé pour déterminer si la covariable C (C_1 ou C_2) induit significativement du DIF sur l'item j. Les hypothèses nulle et alternative de ce test sont les mêmes qu'à l'étape 4, cf. équation (5.5).

– L'algorithme sélectionne ensuite le modèle associé au test le plus significatif après une correction de Bonferroni (risque de première espèce : 5%/nombre de paires candidates). On suppose que la paire associée à ce modèle, celle pour laquelle la contrainte d'invariance a été relâchée, est affectée par du DIF. Cette paire sera dorénavant notée (item j^*, covariable C^*).
Si aucun test n'est significatif, l'algorithme passe alors directement à l'étape 6.
Sinon, la forme du DIF induite par la covariable C^* sur l'item $j*$ est déterminée grâce à un nouveau test statistique réalisé avec un risque de première espèce à 5% et dont les hypothèses nulle et alternative sont :

$$\begin{aligned} &\mathrm{H}_0 : \forall\ p,\ \gamma_{j^*p}^{(C^*)} = \gamma_{j^*}^{(C^*)} \text{ (DIF homogène)} \\ &\mathrm{H}_1 : \exists\ p,\ p' :\ \gamma_{j^*p}^{(C^*)} \neq \gamma_{j^*p'}^{(C^*)} \text{ (DIF non homogène)} \end{aligned} \tag{5.6}$$

– Le modèle C est ensuite mis à jour pour prendre en compte la forme de DIF mise en évidence.
Si le test précédent est significatif, on suppose alors que le DIF est non homogène et les paramètres de DIF $\gamma_{j^*p}^{(C^*)}$ associés à la paire retenue sont librement estimés.
Si le test est non significatif, on considère le DIF homogène et les paramètres de DIF $\gamma_{j^*p}^{(C^*)}$ sont estimés, mais contraints à être constants : $\forall\ p,\ \gamma_{j^*p}^{(C^*)} = \gamma_{j^*}^{(C^*)}$.

La paire retenue ne sera plus testée. L'étape est répétée sur les paires candidates restantes. Elle s'arrête lorsqu'il n'y a plus de paire associée à un test significatif ou lorsque toutes les paires candidates ont été testées ou juste avant de libérer la contrainte d'invariance du dernier item *anchor* pour une covariable donnée (sinon le modèle ne sera plus identifiable).

Étape 6. Estimation de l'effet des covariables sur le niveau moyen de la variable latente

Au cours de cette dernière étape, un modèle final (modèle D) est estimé. Si l'algorithme n'a pas mis évidence de DIF alors le modèle D correspond au modèle B. Sinon, le modèle D correspond à la dernière version du modèle C obtenu à la fin de l'étape 5. Ce modèle permet d'obtenir l'estimation de l'effet des covariables C_1 et C_2 sur le niveau moyen de la variable latente (ajusté sur le DIF éventuellement détecté au cours de l'algorithme).

Notes : C désigne ici indistinctement la covariable C_1 ou C_2. L'ensemble des PCM sur lesquels repose l'algorithme sont estimés par maximum de vraisemblance marginale. Pour tous les modèles, on suppose que les covariables n'influent pas la variance de la variable aléatoire. Les tests statistiques réalisés lors des étapes 4 et 5 sont des tests pour hypothèses jointes.

Dans la suite du manuscrit, cette extension sera désignée sous le nom de *ROSALI-DIF FORWARD*. Cette extension est similaire à la version originelle avec une covariable. On peut néanmoins remarquer l'introduction de l'étape 4 de *screening* qui n'existait pas auparavant. Cette étape vise à déterminer les paires qui seront candidates pour la détection du DIF et celles qui ne le seront pas (les paires dites *anchor*). Elle permet de réaliser un premier tri dans les paires item-covariable et donc de limiter le nombre de tests effectués lors de l'étape 5 (étape itérative). En effet, l'étape 5 implique une succession de tests pour la détection du DIF. Sans *screening*, $2 \times J$ tests seraient réalisés lors de l'itération n°1, $2 \times J - 1$ tests lors de l'itération n°2, etc. La mise en place de l'étape de *screening* limite ce nombre de tests puisqu'il y aura, à chaque itération, autant de tests que de paires candidates (ce nombre étant attendu inférieur à $2 \times J$). Afin d'éviter de détecter du DIF à tort à cause de la multiplicité des tests, le seuil de significativité a été ajusté pour prendre en compte le nombre de tests réalisés. Ainsi, à chaque itération de l'étape 5, les tests sont réalisés avec un risque de première espèce $\alpha = 5\%/\text{Nombre de paires candidates}$.

5.3.2 ROSALI-DIF BACKWARD : 2e version de l'extension

Une version alternative de cette extension a également été explorée, avec la même philosophie, mais où l'étape 5 est basée sur un processus *backward* au lieu d'un processus *forward*. Dans cette version alternative, toutes les contraintes d'invariance associées aux paires candidates sont relâchées simultanément au début de l'étape 5 (au lieu d'une par une). L'effet du DIF est ensuite testé pour chaque paire candidate et les paires associées à des tests non significatifs deviennent toutes des paires *anchor*. L'étape est répétée jusqu'à ce que les paires qui sont encore candidates aient toutes des tests significatifs. Cette version alternative a été nommée *ROSALI-DIF BACKWARD* et est décrite dans les pages suivantes.

ROSALI-DIF BACKWARD : 2eme version de l'extension

Étapes 1 à 4 : Identiques à celles de ROSALI-DIF FORWARD

Étape 5a. Sélection des paires item-covariable affectées par du DIF parmi les paires candidates

Cette étape est itérative. Elle a pour objectif de sélectionner les paires item-covariable affectées par du DIF parmi les paires candidates. Un nouveau modèle, appelé modèle C, est introduit. Le modèle C est basé sur le modèle B (modèle sans DIF) mais où toutes les contraintes d'invariance associées aux paires candidates sont relâchées simultanément. Ainsi, pour les paires *anchor*, les paramètres de DIF $\gamma_{jp}^{(C)}$ de l'équation (5.4) sont contraints à 0 (pas de DIF). En revanche, pour toutes les paires candidates, les paramètres de DIF sont librement estimés.

À partir du modèle C, un test statistique est effectué pour chaque paire candidate afin de déterminer si l'effet DIF induit par la covariable C (C_1 ou C_2) sur l'item j est significatif ou non. Les hypothèses nulle et alternative sont les mêmes qu'à l'étape 4, c'est-à-dire :

$$\begin{aligned} &\mathrm{H}_0 : \forall\ p, \gamma_{jp}^{(C)} = 0 \quad \text{(La covariable } C \text{ n'induit pas de DIF sur l'item } j\text{)} \\ &\mathrm{H}_1 : \exists\ p : \gamma_{jp}^{(C)} \neq 0 \quad \text{(La covariable } C \text{ induit du DIF sur l'item } j\text{)} \end{aligned} \tag{5.5}$$

Si toutes les paires candidates sont associées à un test significatif après une correction de Bonferroni (niveau de signification = 5%/nombre de paires candidates), on suppose la présence de DIF pour toutes ces paires et l'algorithme passe à l'étape suivante. Dans le cas contraire, des contraintes d'invariance sont ajoutées pour les paires associées aux tests non significatifs et le modèle C est mis à jour avec ces nouvelles contraintes (ces paires deviennent donc des paires *anchor* et ne seront plus considérées comme candidates pour la détection du DIF). L'étape 5 est répétée jusqu'à ce que les paires encore candidates aient toutes des tests significatifs ou jusqu'à ce qu'il n'y ait plus aucune paire candidate (pas de DIF détecté).

À la fin de cette étape, si aucun item *anchor* n'est identifié pour l'une des covariables, l'algorithme est alors interrompu (l'effet de la covariable en question sur le niveau moyen de la variable latente ne pouvant être estimé).

Étape 5b : Évaluation de la forme du DIF

Pour chacune des paires sélectionnées à l'étape 5a (paires affectées par du DIF), un test statistique est réalisé à partir des dernières estimations du modèle C pour déterminer la forme du DIF (risque de première espèce à 5%). Les hypothèses nulle et alternative sont :

$$\begin{aligned} &\mathrm{H}_0 : \forall\ p,\ \gamma_{jp}^{(C)} = \gamma_{j}^{(C)}\ \text{(DIF homogène)} \\ &\mathrm{H}_1 : \exists\ p,\ p' :\ \gamma_{jp}^{(C)} \neq \gamma_{jp'}^{(C)}\ \text{(DIF non homogène)} \end{aligned} \tag{5.6}$$

Une fois toutes les paires affectées par du DIF testées, le modèle C est mis à jour pour prendre en compte la ou les forme(s) de DIF mise(s) en évidence. Pour les paires associées à un test significatif, les paramètres de DIF $\gamma_{jp}^{(C)}$ associés seront librement estimés (DIF non homogène). Pour les paires associées à un test non significatif, les paramètres de DIF seront estimés, mais contraints à être égaux (DIF homogène) : $\forall\ p,\ \gamma_{jp}^{(C)} = \gamma_{j}^{(C)}$.

Étape 6. Estimation de l'effet des covariables sur le niveau moyen de la variable latente

Au cours de cette dernière étape, un modèle final (modèle D) est estimé. Si l'algorithme n'a pas mis évidence de DIF alors le modèle D correspond au modèle B. Sinon, le modèle D correspond à la dernière version du modèle C obtenu à la fin de l'étape 5b. Ce modèle permet d'obtenir l'estimation de l'effet des covariables C_1 et C_2 sur le niveau moyen de la variable latente (ajusté sur le DIF éventuellement détecté au cours de l'algorithme).

Notes : C désigne ici indistinctement la covariable C_1 ou C_2.
L'ensemble des PCM sur lesquels repose l'algorithme sont estimés par maximum de vraisemblance marginale. Pour tous les modèles, on suppose que les covariables n'influent pas la variance de la variable latente. Les tests statistiques réalisés lors des étapes 4 et 5 (a et b) sont des tests pour hypothèses jointes. Au cours de l'étape 5a, l'effet d'une covariable sur le niveau moyen de la variable latente est estimé dès qu'un item anchor est identifié pour cette covariable.

Les deux algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD sont représentés conjointement dans la figure 5.2.

Les performances de ces deux algorithmes pour détecter du DIF ont été évaluées à l'aide d'une étude de simulation présentée dans la section 5.5. En parallèle, une procédure basée sur l'estimation pénalisée d'un PCM a également été évaluée dans les mêmes conditions. Cette procédure est décrite danse la section suivante avec deux covariables binaires.

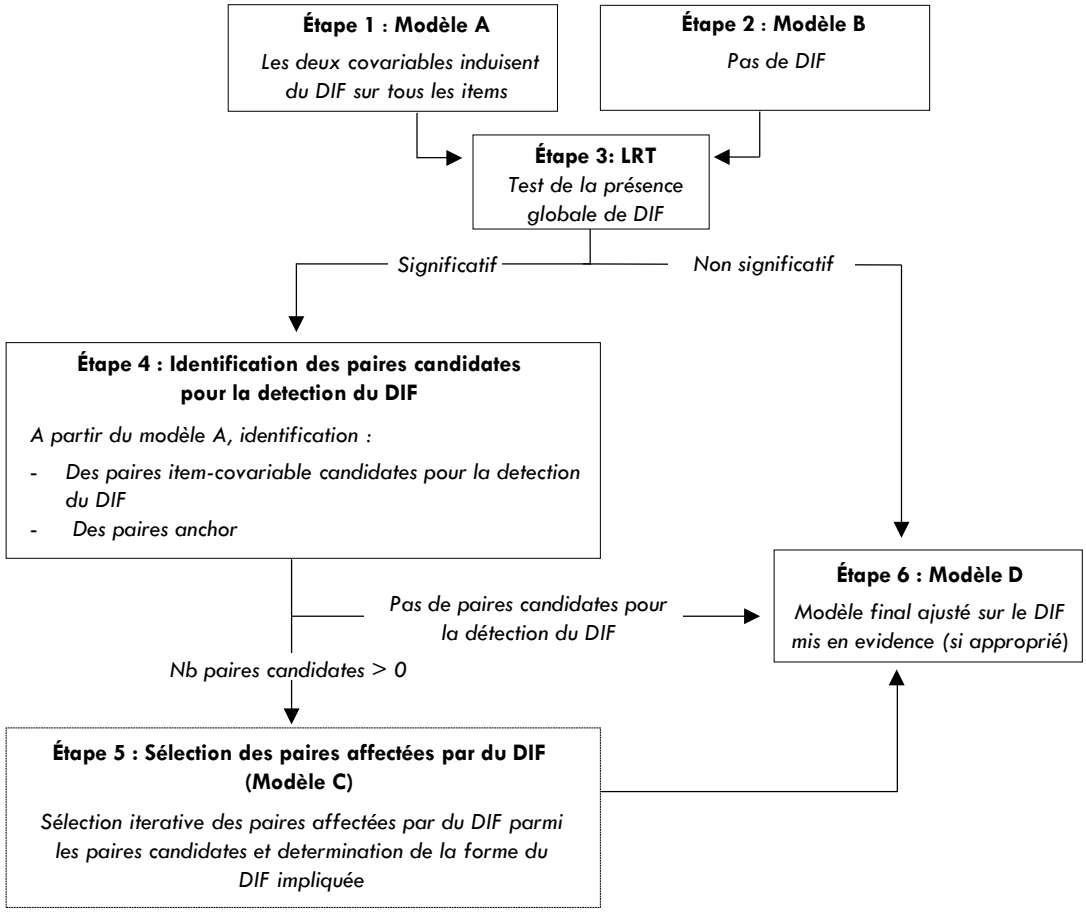


FIGURE 5.2 – Représentation graphique des algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD

Notes : Nb : Nombre, LRT : test du rapport de vraisemblance (likelihood-ratio test)

5.4 Détection du DIF par pénalisation d'un PCM

Une méthode de détection du DIF en présence de plusieurs covariables et reposant sur la pénalisation de la vraisemblance d'un PCM ou d'un PCM généralisé (GPCM) a été décrite de façon détaillée par Schauberger et Mair [19]. Bien que permettant d'estimer un GPCM, cette méthode ne s'intéresse qu'à des différences entre groupes sur les paramètres de seuil des items. En utilisant la pénalisation de la vraisemblance, les auteurs ont traduit la détection du DIF en un problème de sélection de paramètres à estimer, dont l'objectif est de répondre à la question : "Quels paramètres DIF valent la peine d'être estimés ?". Ces auteurs ont en fait proposé deux méthodes présentées dans les pages suivantes dans le cadre des modèles de la famille de Rasch (PCM).

5.4.1 Méthode n°1 : Détection du DIF homogène

La première méthode proposée par Schauberger et Mair fait l'hypothèse forte que le DIF est homogène, c'est-à-dire que les différences de paramètres de seuil induites par une covariable C ont le même sens et la même magnitude pour tous les paramètres de seuil associés à un item. Mathématiquement, ces auteurs supposent donc que lorsqu'un item j est affecté par du DIF induit par une covariable C, les paramètres de DIF $\gamma_{jp}^{(C)}$ sont constants : $\forall\ p,\ \gamma_{jp}^{(C)} = \gamma_{j}^{(C)}$.

Ainsi, si l'on considère deux covariables binaires C_1 et C_2 suspectées d'induire du DIF, le PCM à estimer de façon pénalisée s'écrit alors :

$$P\left(X_{ij} = x \mid \theta_i, \delta_{j1}, \ldots, \delta_{jM_j-1}, C_{1_i}, C_{2_i}, \beta_1, \beta_2, \gamma_j^{(C_1)}, \gamma_j^{(C_2)}\right) = \frac{\exp\left(x\left[\theta_i + \beta_1 C_{1_i} + \beta_2 C_{2_i}\right] - \sum_{p=1}^{x}\left[\delta_{jp} + \gamma_j^{(C_1)} C_{1_i} + \gamma_j^{(C_2)} C_{2_i}\right]\right)}{\sum_{l=0}^{M_j-1}\exp\left(l\left[\theta_i + \beta_1 C_{1_i} + \beta_2 C_{2_i}\right] - \sum_{p=1}^{l}\left[\delta_{jp} + \gamma_j^{(C_1)} C_{1_i} + \gamma_j^{(C_2)} C_{2_i}\right]\right)} \tag{5.7}$$

Pour estimer ce modèle avec une pénalisation de type lasso, Schauberger et Mair ont cherché à maximiser la log-vraisemblance ℓ du modèle, pénalisée par la somme pondérée des valeurs absolues des paramètres de DIF [19] :

$$\ell_{\text{pénalisée}} = \ell - \lambda \times \sum_{j=1}^{\text{Nb d'items}} \sum_{k=1}^{2} w_{jk} |\gamma_j^{(C_k)}| \tag{5.8}$$

Dans cette équation, λ désigne le paramètre de pénalisation (*tuning parameter*) et $\sum_j \sum_k w_{jk} |\gamma_j^{(C_k)}|$ est le terme de pénalisation de type lasso proposé par les auteurs. Dans ce terme de pénalisation, w_{jk} est un poids adaptatif correspondant à l'inverse de la valeur absolue de l'estimation du paramètre $\gamma_j^{(C_k)}$ obtenue en estimant le modèle avec une pénalisation de Ridge [1] : $w_{jk} = 1/|\widehat{\gamma_j^{(C_k)}}^{Ridge}|$. Schauberger et Mair ont choisi d'utiliser des poids adaptatifs pour que les paramètres associés à de faibles estimations avec Ridge soient plus sévèrement pénalisés que les paramètres avec de grandes estimations.

On peut remarquer dans l'équation (5.8) que la pénalisation ne s'applique qu'aux paramètres de DIF $\gamma_j^{(C_k)}$. L'estimation de chacun de ces paramètres a donc un "coût" qui doit être contre-balancé par un gain de vraisemblance pour qu'il vaille effectivement la peine d'être estimé. Les paramètres de distribution de la variable latente sont en revanche censés être toujours estimés (sans pénalisation) [19].

Lors de la maximisation de la vraisemblance pénalisée donnée par l'équation (5.8), le nombre de paramètres de DIF estimés non nuls dépend entièrement du choix du paramètre de pénalisation λ (le paramètre qui contrôle la force de la pénalisation). Lorsque ce paramètre est égal à zéro, tous les paramètres sont estimés : il n'y a pas de pénalisation. Au contraire, aucun paramètre n'est estimé lorsque ce paramètre tend vers $+\infty$.

1. Les auteurs ne fournissent pas d'informations complémentaires sur la méthode employée pour l'obtention de ces poids adaptatifs. Ils indiquent uniquement qu'ils se sont basés sur les valeurs des paramètres obtenues en estimant le modèle avec une pénalisation de Ridge (il s'agit d'une autre méthode d'estimation régularisée où la log-vraisemblance du modèle est pénalisée par la somme des carrés des paramètres de DIF).

En pratique, afin de sélectionner le paramètre de pénalisation adéquat, une large grille de valeurs à explorer est d'abord identifiée : $\lambda_{min}, ..., \lambda_{max}$. Pour chaque valeur λ de cette grille, un PCM est estimé en maximisant la log-vraisemblance pénalisée associée à cette valeur. Une fois que tous les PCM associés aux valeurs de pénalisation de la grille ont été estimés, les estimations du paramètre de DIF γ_{jp} de chaque paire item-covariable sont tracées en fonction des différentes valeurs du paramètre λ. Ce graphique est appelé un *parameter path*. Un exemple fictif est donné dans la figure 5.3 pour une paire item-covariable quelconque.

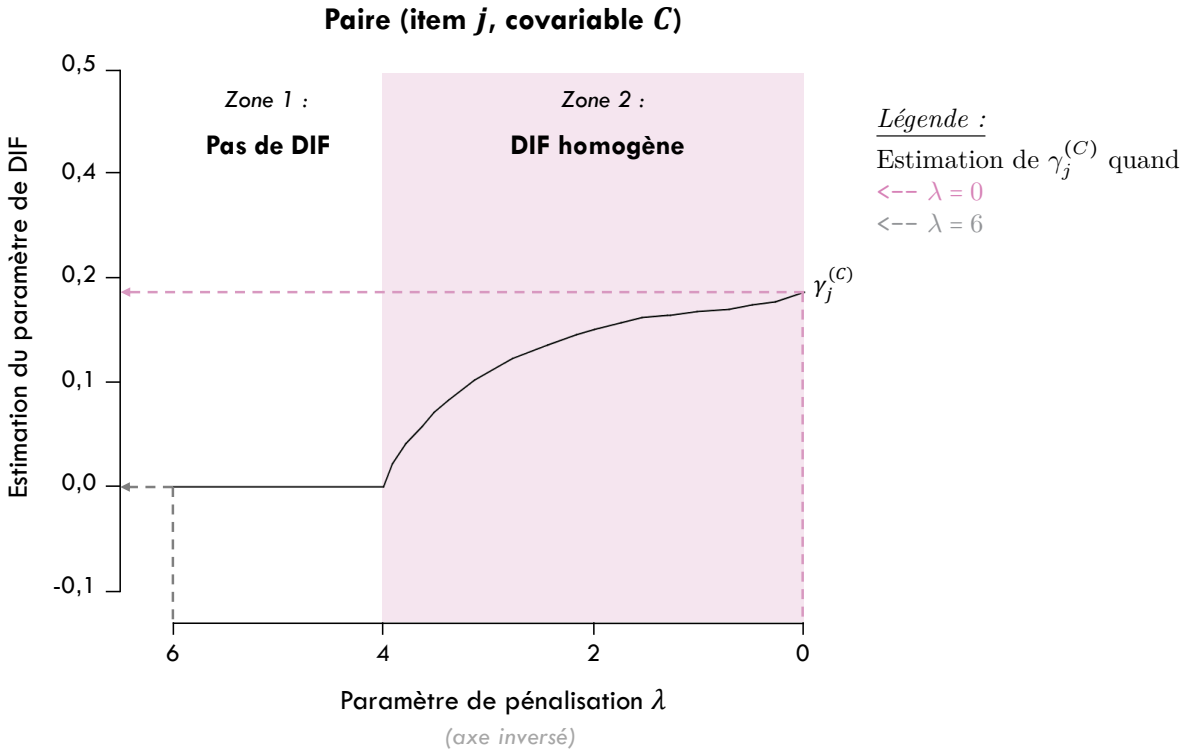


FIGURE 5.3 – Exemple fictif de *parameter path* pour le paramètre de DIF $\gamma_j^{(C)}$ avec la méthode de pénalisation n°1

Notes :

Dans ce graphique, la grille de valeurs explorées pour le paramètre de pénalisation va de 0 à 6. Pour les modèles associés à un paramètre de pénalisation compris entre 4 et 6 (zone 1), le paramètre de DIF $\gamma_j^{(C)}$ associé à la paire (item j, covariable C) est fixé à 0 par la pénalisation. Ainsi, si l'un de ces modèles était choisi, la procédure ne mettrait pas en évidence de DIF sur cette paire. En revanche, le parameter path associé à $\gamma_j^{(C)}$ s'écarte de 0 dès que le paramètre de pénalisation λ est inférieur à 4 (zone 2). Ainsi, si l'un des modèles associés à un paramètre de pénalisation $\lambda < 4$ était choisi, du DIF homogène serait mis en évidence sur la paire.

Le paramètre de pénalisation retenu est celui associé au modèle minimisant le critère d'information bayésien (BIC), défini ici par :

$$BIC(\lambda) = -2\ell_\lambda(\cdot) + df(\lambda)log(N) \tag{5.9}$$

Où $\ell_\lambda(\cdot)$ désigne la log-vraisemblance du modèle obtenu avec le paramètre de pénalisation λ, $df(\lambda)$ est le nombre total de paramètres estimés non nuls dans ce modèle et N est l'effectif de l'échantillon. Le BIC a été choisi par les auteurs, car il s'agit d'un critère usuel de sélection des paramètres qui permet d'appliquer une sélection plus conservatrice que le critère d'information d'Akaike (Akaike Information Criterion, AIC) [19]. Les résultats de la détection du DIF s'obtiennent ensuite à partir du modèle minimisant le BIC. Pour chaque paire (item j, covariable C), on conclura que la covariable C (C_1 ou C_2) induit du DIF sur l'item j si et seulement si le paramètre de DIF associé à la paire est non nul. Comme vu précédemment, on conclura forcément à du DIF homogène avec cette méthode.

Schauberger et Mair indiquent dans leur manuscrit que tant que la pénalisation est assez forte, aucun problème d'identifiabilité du modèle ne devrait être rencontré (la pénalisation assurant qu'au moins un paramètre de DIF soit nul pour chacune des covariables introduites dans l'analyse). Ils ont évalué cette méthode de détection par simulations dans le cas d'un échantillon composé de 500 individus répondant à un questionnaire contenant 20 items, dont 3 ou 4 étaient affectés par du DIF homogène induit par des covariables distinctes [19].

5.4.2 Méthode n°2 : Détection du DIF sans présumer de sa forme

La seconde méthode proposée par Schauberger et Mair [19] est plus générale et ne fait pas l'hypothèse que le DIF est nécessairement homogène. Avec deux covariables binaires C_1 et C_2 suspectées d'induire du DIF, le modèle PCM à estimer de façon pénalisé s'écrit alors comme décrit précédemment dans l'équation (5.4) :

$$P\left(X_{ij} = x \mid \theta_i, \delta_{j1}, \ldots, \delta_{jM_j-1}, C_{1_i}, C_{2_i}, \beta_1, \beta_2, \gamma_{j1}^{(C_1)}, \ldots, \gamma_{jM_j-1}^{(C_1)}, \gamma_{j1}^{(C_2)}, \ldots, \gamma_{jM_j-1}^{(C_2)}\right) = \frac{\exp\left(x\left[\theta_i + \beta_1 C_{1_i} + \beta_2 C_{2_i}\right] - \sum_{p=1}^{x}\left[\delta_{jp} + \gamma_{jp}^{(C_1)} C_{1_i} + \gamma_{jp}^{(C_2)} C_{2_i}\right]\right)}{\sum_{l=0}^{M_j-1} \exp\left(l\left[\theta_i + \beta_1 C_{1_i} + \beta_2 C_{2_i}\right] - \sum_{p=1}^{l}\left[\delta_{jp} + \gamma_{jp}^{(C_1)} C_{1_i} + \gamma_{jp}^{(C_2)} C_{2_i}\right]\right)}$$

Ici encore, les auteurs ne cherchaient pas à maximiser la vraisemblance de ce modèle, mais la log-vraisemblance pénalisée. Néanmoins, la forme du terme de pénalisation a été modifiée, avec l'ajout d'un second terme dont l'objectif est de déterminer pour chaque paire (item j, covariable C) s'il vaut mieux estimer des paramètres de DIF égaux (DIF homogène) ou non (DIF non homogène). La log-vraisemblance pénalisée associée à ce modèle s'écrit :

$$\begin{aligned}\ell_{\text{pénalisée}} = \ell - \lambda \times \Bigg(& \sum_{j=1}^{\text{Nb d'items}} \sum_{k=1}^{2} \sum_{p=1}^{M_j-1} w_{jk(p)} \left|\gamma_{jp}^{(C_k)}\right| \\ & + \sum_{j=1}^{\text{Nb d'items}} \sum_{k=1}^{2} \sum_{p\,<\,p'} w_{jk(p,\,p')} \left|\gamma_{jp}^{(C_k)} - \gamma_{jp'}^{(C_k)}\right| \Bigg)\end{aligned} \tag{5.10}$$

Les poids $w_{jk(p)}$ correspondent à l'inverse de la valeur absolue de l'estimation de Ridge du paramètre $w_{jk(p)} = 1/\left|\widehat{\gamma_{jp}^{(C_k)}}^{Ridge}\right|$ et les poids $w_{jk(p,\,p')}$ sont définis par : $w_{jk(p,\,p')} = 1/\left|\widehat{\gamma_{jp}^{(C_k)}}^{Ridge} - \widehat{\gamma_{jp'}^{(C_k)}}^{Ridge}\right|$. Comme précédemment, les poids $w_{jk(p)}$ permettent de pénaliser plus sévèrement les paramètres associés à de faibles estimations de Ridge. Les poids $w_{jk(p,\,p')}$ permettent quant à eux de pénaliser plus sévèrement des écarts faibles entre les paramètres de DIF d'une même paire $\gamma_{jp}^{(C_k)}$ et $\gamma_{jp'}^{(C_k)}$.

De façon similaire à la méthode n°1, plusieurs PCM sont estimés (autant de modèles que de paramètres de pénalisation dans la grille de valeurs considérées). Les estimations de ces modèles sont résumées dans les DIF *parameters path* et le modèle sélectionné est celui ayant le plus petit BIC (cf. équation (5.9)). Avec cette méthode, deux situations principales peuvent se présenter lors de la recherche de DIF, comme illustré avec la figure 5.4 :

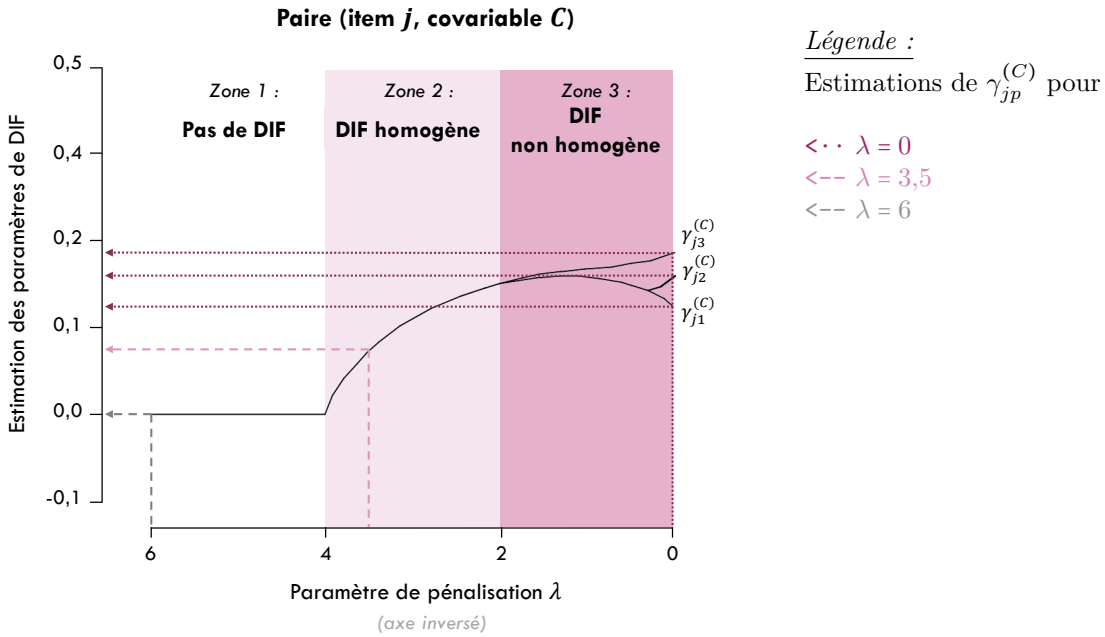
- Soit le graphique présente une zone où les paramètres de DIF $\gamma_{jp}^{(C)}$ sont estimés, mais contraints à être égaux (zone 2 du graphique a) ;
- Soit le graphique ne présente pas une telle zone (graphique b).

Les résultats de la détection du DIF s'obtiennent ensuite à partir du modèle minimisant le BIC. Pour chaque paire (item j, covariable C), on conclura que la covariable C (C_1 ou C_2) induit du DIF sur l'item j si et seulement si l'un des paramètres de DIF associé à la paire est non nul dans le modèle retenu : $\exists\ p\ : \gamma_{jp}^{(C)} \neq 0$. Pour une paire donnée, on conclura à du DIF homogène si les paramètres de DIF associés sont estimés non nuls mais contraints à être égaux par la pénalisation. On conclura à du DIF non homogène dans les autres cas.

À notre connaissance, cette méthode n'a jamais été évaluée par simulations, ces performances pour la détection du DIF ne sont donc pour l'heure pas connues. Nous avons donc choisi de l'évaluer en parallèle des deux algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD afin de : (i) fournir des données empiriques sur cette méthode et (ii) avoir une autre méthode à laquelle comparer les deux versions de l'extension de la partie 1 de ROSALI. Nous avons sélectionné cette méthode en particulier puisqu'elle était basée sur une philosophie très différente de celle de ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD.

Dans la suite du manuscrit, cette méthode de pénalisation sera appelée PCM-Lasso (estimation pénalisée par lasso d'un PCM). Il est à noter que dans l'implémentation actuelle de cette méthode, un paramètre de discrimination commun à tous les items est estimé. Ce paramètre est égal à 1 dans les deux algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD.

a. Présence d'une zone où les paramètres de DIF sont estimés, mais contraints à être égaux



b. Absence de zone où les paramètres de DIF sont estimés mais contraints à être égaux

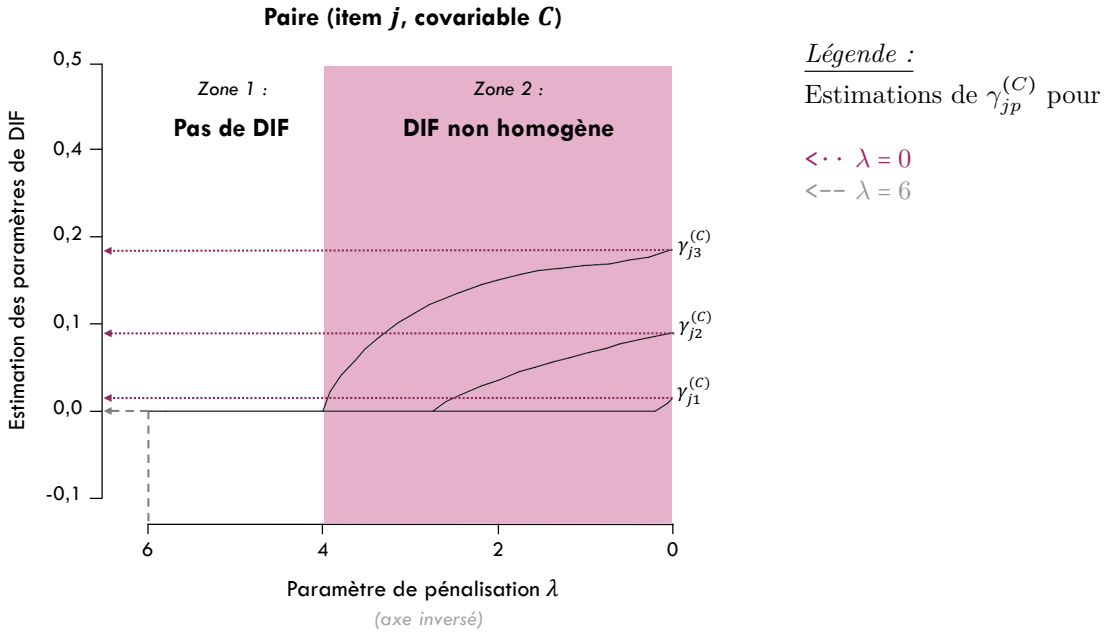


FIGURE 5.4 – Exemple fictif de *parameter path* pour les paramètres DIF $\gamma_{jp}^{(C)}$ avec la méthode de pénalisation n°2

Notes :

Dans ces deux graphiques, la grille de valeurs explorées pour le paramètre de pénalisation va de 0 à 6.

Graphique a *– Pour les modèles associés à un paramètre de pénalisation compris entre 4 et 6 (zone 1), tous les paramètres de DIF $\gamma_{jp}^{(C)}$ associés à la paire (item j, covariable C) sont fixés à 0 par la pénalisation. Ainsi, si l'un de ces modèles était choisi, la procédure ne mettrait pas en évidence de DIF sur cette paire. Pour les modèles associés à un paramètre de pénalisation compris entre 2 et 4 les parameter paths associés à $\gamma_{j1}^{(C)}$, $\gamma_{j2}^{(C)}$ et $\gamma_{j3}^{(C)}$ s'écartent de 0 de façon simultanée (les paramètres sont contraints par la pénalisation à être égaux). Ainsi, si l'un de ces modèles était choisi, du DIF homogène serait mis en évidence. Enfin, pour les modèles associés à un paramètre de pénalisation compris entre 0 et 2, les parameter paths des trois paramètres se séparent (ils ne sont plus contraints à être égaux par la pénalisation). Du DIF non homogène serait donc mis en évidence sur la paire si on sélectionnait l'un de ces modèles.*

Graphique b *– Pour les modèles associés à un paramètre de pénalisation compris entre 4 et 6 (zone 1), tous les paramètres de DIF $\gamma_{jp}^{(C)}$ associés à la paire (item j, covariable C) sont fixés à 0 par la pénalisation. Ainsi, si l'un de ces modèles était choisi, la procédure ne mettrait pas en évidence de DIF sur cette paire. En revanche, le parameter path associé à $\gamma_{j3}^{(C)}$ s'écarte de 0 dès que le paramètre de pénalisation λ est inférieur à 4 (zone 2). Ainsi, si l'un des modèles associés à un paramètre de pénalisation $\lambda < 4$ était choisi, du DIF non homogène serait mis en évidence sur la paire.*

5.5 Étude de simulation

L'intérêt des études de simulation a déjà été rappelé dans le chapitre 4, section 4.3.

L'objectif de l'étude de simulation que nous avons réalisée ici était d'évaluer les performances des algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD et les comparer à celles obtenues en utilisant la méthode PCM-Lasso. Nous avons étudié des situations variées qui incluaient un nombre modéré d'items polytomiques (ce qui est usuel avec les données rapportées par les patients), deux covariables binaires éventuellement corrélées et suspectées d'induire différentes formes de DIF.

5.5.1 Simulation des données

Lors de cette étude, nous avons simulé les réponses de $N = 400$ ou 800 individus à un questionnaire unidimensionnel composé de $J = 4$ ou 7 items polytomiques (item 1, item 2, ... , item J). Il y avait pour chaque item $M = 4$ modalités de réponse qui étaient numérotées de 0 à 3. Les niveaux individuels de la variable latente ont été tirés d'une loi normale centrée-réduite $\mathcal{N}(0,1)$ et les réponses au questionnaire ont été générées selon un PCM. Les paramètres de seuil des items δ_{jp} ont été choisis pour : (i) être centrés sur 0 (la moyenne de la variable latente) et (ii) couvrir tout le continuum de la variable latente. Plus spécifiquement, pour chaque item j, le premier paramètre de seuil δ_{j1} (associé à la modalité de réponse "1") a d'abord été initialisé au $\frac{j}{j+1}$ième quantile de la loi normale centrée réduite. Les paramètres de seuil associés aux modalités de réponse p suivantes ($p > 1$) ont été choisis pour être régulièrement espacés à partir du premier paramètre δ_{j1} et de façon à ce que l'écart entre le premier et le dernier paramètre de seuil de l'item j soit égal à 2 : $\delta_{jp} = \delta_{j1} + (p-1) \times \frac{2}{M-2},\ p = 2, \ldots, M-1$. Pour terminer, les paramètres de seuil des items ont été centrés sur 0 en soustrayant la moyenne des paramètres $\bar{\delta} = \frac{\sum_{j,p} \delta_{jp}}{J \times (M-1)}$. Les valeurs des paramètres de seuil des items sont données dans le tableau 5.1 pour chaque item. Elles sont également représentées graphiquement dans la figure 5.5.

TABLEAU 5.1 – Paramètres de seuil utilisés pour l'étude de simulation

	δ_{j1}	δ_{j2}	δ_{j3}
Pour $J = 4$ items			
Item 1	-1,84	-0,84	0,16
Item 2	-1,25	-0,25	0,75
Item 3	-0,75	0,25	1,25
Item 4	-0,16	0,84	1,84
Pour $J = 7$ items			
Item 1	-2,15	-1,15	-0,15
Item 2	-1,67	-0,67	0,33
Item 3	-1,32	-0,32	0,68
Item 4	-1,00	0,00	1,00
Item 5	-0,68	0,32	1,32
Item 6	-0,33	0,67	1,67
Item 7	0,15	1,15	2,15

Notes : J : nombre d'items
δ_{jp} : paramètre de seuil associé à la modalité de réponse p de l'item j

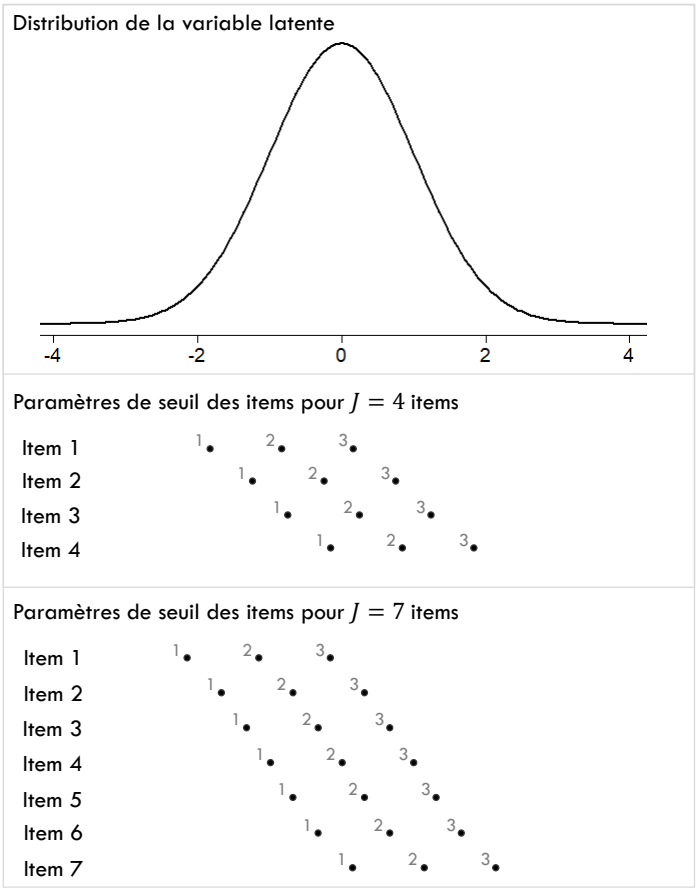


FIGURE 5.5 – Représentation graphique des paramètres de seuil des items

5.5.2 Opérationnalisation du DIF

Pour cette étude de simulation, nous souhaitions évaluer les performances de trois méthodes de détection du DIF en présence de deux covariables binaires C_1 et C_2 éventuellement corrélées. Nous nous sommes placés dans le cas simple où les effectifs de chaque covariable étaient bien équilibrés comme présenté dans le tableau 5.2.

TABLEAU 5.2 – Description de l'échantillon considéré dans l'étude de simulation

	Échantillon total (N = 400 ou 800)
Covariable C_1	
Modalité 0	50%
Modalité 1	50%
Covariable C_2	
Modalité 0	50%
Modalité 1	50%

Nous avons étudié deux grands cas de figure au niveau de ces deux covariables :

– Le cas où les deux covariables sont indépendantes ;

– Le cas où les deux covariables sont corrélées.

Dans le cas où les covariables étaient indépendantes, nous avons simulé des groupes de façon à obtenir :

- 25% d'individus $C_1 = 0$ et $C_2 = 0$
- 25% d'individus $C_1 = 1$ et $C_2 = 0$
- 25% d'individus $C_1 = 1$ et $C_2 = 1$
- 25% d'individus $C_1 = 1$ et $C_2 = 1$

Tableau de contingence

		Covariable C1		
		0	1	Total
Covariable C_2	0	$N/4$	$N/4$	$N/2$
	1	$N/4$	$N/4$	$N/2$
	Total	$N/2$	$N/2$	N

Dans le cas où les covariables étaient corrélées, nous avons choisi de déséquilibrer les effectifs du tableau de contingence entre les deux covariables C_1 et C_2 de sorte qu'il y ait :

- 37,5% d'individus $C_1 = 0$ et $C_2 = 0$
- 12,5% d'individus $C_1 = 1$ et $C_2 = 0$
- 12,5% d'individus $C_1 = 0$ et $C_2 = 1$
- 37,5% d'individus $C_1 = 1$ et $C_2 = 1$

Tableau de contingence

		Covariable C1		
		0	1	Total
Covariable C_2	0	$3N/8$	$N/8$	$N/2$
	1	$N/8$	$3N/8$	$N/2$
Total		$N/2$	$N/2$	N

Ces deux grands cas de figure sont représentés dans la figure 5.6. Le graphique (a) représente le cas où les deux covariables ne sont pas corrélées et le graphique (b) illustre le cas où elles le sont.

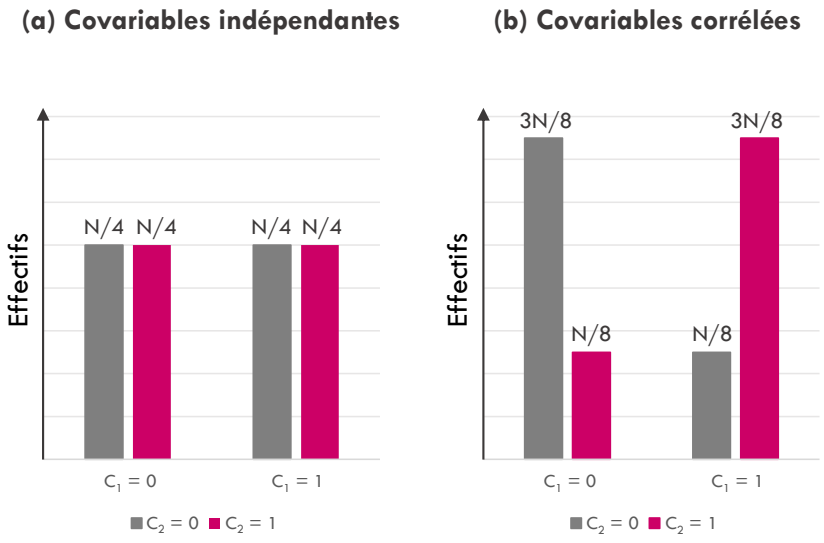


FIGURE 5.6 – Diagramme en barres croisant les covariables C_1 et C_2

Pour opérationnaliser un fonctionnement différentiel de l'item j induit par la covariable C (C_1 ou C_2), nous avons décalé les paramètres de seuil de cet item entre les groupes formés par les covariables. Ce décalage est représenté mathématiquement par les paramètres $\gamma_{jp}^{(C)}$ de l'équation 5.4 (paramètres de DIF). Trois configurations pour le DIF ont été considérées :

Configuration n°1 : Les deux covariables n'étaient pas corrélées et elles induisaient toutes deux du DIF sur un item différent.

Plus précisément :

- Quand le questionnaire était composé de 4 items, la covariable C_1 induisait du DIF sur l'item 2 et la covariable C_2 induisait du DIF sur l'item 3.
- Quand le questionnaire était composé de 7 items, la covariable C_1 induisait du DIF sur l'item 3 et la covariable C_2 induisait du DIF sur l'item 5.

Configuration n°2 : Les deux covariables n'étaient pas corrélées et elles induisaient toutes deux du DIF sur le même item.

Plus précisément :

- Quand le questionnaire était composé de 4 items, les deux covariables induisaient du DIF sur l'item 2.
- Quand le questionnaire était composé de 7 items, les deux covariables induisaient du DIF sur l'item 3.

Configuration n°3 : Les deux covariables étaient corrélées et seule la covariable C_1 induisait du DIF sur deux items.

Plus précisément :

- Quand le questionnaire était composé de 4 items, la covariable C_1 induisait du DIF sur les items 2 et 3.
- Quand le questionnaire était composé de 7 items, la covariable C_1 induisait du DIF sur les items 3 et 5.

Ainsi, pour la configuration n°1, les paramètres de seuil des items affectés par du DIF induit par la covariable C (C_1 ou C_2) sont donnés par :

$$\delta_{jp} + \gamma_{jp}^{(C)} \times C = \begin{cases} \delta_{jp} \text{ dans le goupe } C = 0 \\ \delta_{jp} + \gamma_{jp}^{(C)} \text{ dans le groupe } C = 1 \end{cases} \tag{5.11}$$

Pour la configuration n°2, la formule est différente, car les deux covariables induisent du DIF sur le même item. Les paramètres de seuil de l'item affecté par du DIF sont alors donnés par :

$$\delta_{jp} + \gamma_{jp}^{(C_1)} \times C_1 + \gamma_{jp}^{(C_2)} \times C_2 = \begin{cases} \delta_{jp} \text{ dans le groupe } C_1 = 0 \wedge C_2 = 0 \\ \delta_{jp} + \gamma_{jp}^{(C_1)} \text{ dans le groupe } C_1 = 1 \wedge C_2 = 0 \\ \delta_{jp} + \gamma_{jp}^{(C_2)} \text{ dans le groupe } C_1 = 0 \wedge C_2 = 1 \\ \delta_{jp} + \gamma_{jp}^{(C_1)} + \gamma_{jp}^{(C_2)} \text{ dans le groupe } C_1 = 1 \wedge C_2 = 1 \end{cases} \tag{5.12}$$

Enfin, pour la configuration n°3, les paramètres de seuil des items affectés par du DIF induit par la covariable C_1 (la seule à induire du DIF) sont donnés par :

$$\delta_{jp} + \gamma_{jp}^{(C_1)} \times C_1 = \begin{cases} \delta_{jp} \text{ dans le goupe } C_1 = 0 \\ \delta_{jp} + \gamma_{jp}^{(C)} \text{ dans le groupe } C_1 = 1 \end{cases} \tag{5.13}$$

Pour chacune de ces trois configurations, deux formes différentes de DIF ont été explorées (voir figure 5.7) :

- **DIF homogène :** les différences entre groupes dans les paramètres de seuil de l'item vont dans le même sens et ont la même magnitude (les paramètres de DIF sont constants).
- **DIF non homogène :** les différences entre groupes dans les paramètres de seuil de l'item varient en direction et/ou en magnitude. Cette étude de simulation est limitée au cas où les différences augmentent, mais conservent la même direction (les paramètres de DIF ne sont pas constants, mais sont de même signe).

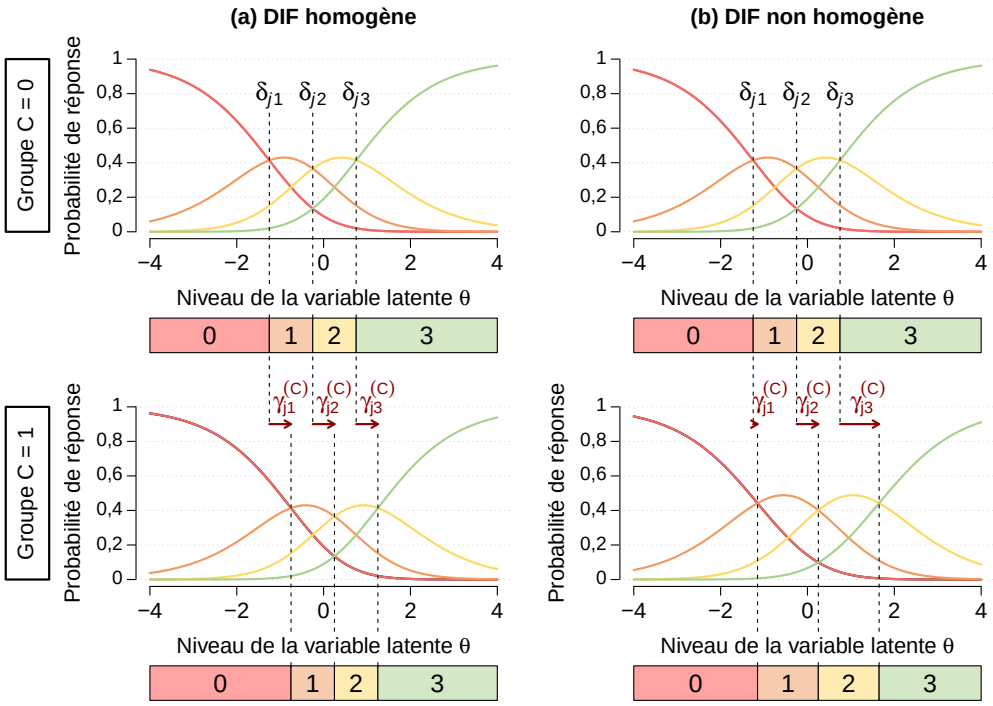


FIGURE 5.7 – Courbes caractéristiques des modalités de réponse d'un item j affecté par du DIF homogène (graphique a) ou non homogène (graphique b) induit par la covariable C

Notes : C désigne indistinctement C_1 ou C_2.
Le décalage des paramètres d'items entre les deux groupes est représenté par les flèches rouges.
Lorsque le DIF est homogène, ces flèches vont dans la même direction et ont la même taille. Lorsque le DIF est non homogène, nous avons choisi de faire uniquement varier la taille des flèches (elles conservent la même direction).

Nous avons également fait varier la taille (ou la magnitude) du DIF, qui pouvait être faible ou moyenne. Les différences entre groupes dans les paramètres de seuil des items affectés par du DIF était d'en moyenne +0.3 pour du DIF de taille faible (respectivement +0.5 pour du DIF de taille moyenne).

Au sein d'un même scénario, le DIF a été simulé selon la règle suivante : il n'y avait qu'une seule forme et qu'une seule taille de DIF par scénario. Ainsi, au sein d'un même scénario, le DIF était soit :

- Homogène et de taille faible ;
- Homogène et de taille moyenne ;
- Non homogène et de taille faible ;
- Non homogène et de taille moyenne.

L'étude de simulation est résumée dans le tableau 5.3. La combinaison de tous ces paramètres de simulation a conduit à un total de 48 scénarios auxquels ont été ajoutés 8 scénarios où aucun DIF n'a été simulé. Tous les scénarios ont été répliqués 500 fois et chaque jeu de données a été analysé avec les trois méthodes de détection du DIF (c.-à-d., les deux algorithmes ROSALI-DIF et la méthode basée sur la pénalisation de la vraisemblance PCM-Lasso).

TABLEAU 5.3 – Résumé de l'étude de simulation

Structure du questionnaire et échantillon	
Nombre d'items	$J = 4, 7$ items
Nombre de modalités de réponse	$M = 4$ modalités de réponse
Taille de l'échantillon	$N = 400$; 800 individus simulés
Variable latente (Θ)	
Moyenne μ, Variance σ^2	$\mu = 0$, $\sigma^2 = 1$
Effet des covariables sur la variable latente	Pas d'effet de la covariable C_1 : $\beta_1 = 0$ Pas d'effet de la covariable C_2 : $\beta_2 = 0$
Covariables à l'origine du DIF	
Configuration n°1	Les covariables C_1 et C_2 ne sont pas corrélées Elles induisent toutes deux du DIF sur un item différent
Configuration n°2	Les covariables C_1 et C_2 ne sont pas corrélées Elles induisent toutes deux du DIF sur le même item
Configuration n°3	Les covariables C_1 et C_2 sont corrélées Seule la covariable C_1 induit du DIF sur deux items
Items affectés par du DIF	
***J* = 4 items**	
Configuration n°1	C_1 induit du DIF sur l'item 2 et C_2 induit du DIF sur l'item 3
Configuration n°2	C_1 et C_2 induisent du DIF sur l'item 2
Configuration n°3	C_1 induit du DIF sur les items 2 et 3
***J* = 7 items**	
Configuration n°1	C_1 induit du DIF sur l'item 3 et C_2 induit du DIF sur l'item 5
Configuration n°2	C_1 et C_2 induisent du DIF sur l'item 3
Configuration n°3	C_1 induit du DIF sur les items 3 et 5
Forme du DIF	
Homogène	L'effet de la covariable sur les paramètres de seuil des items est constant (même direction et magnitude) : $\forall\ p,\ \gamma_{jp}^{(C)} = \gamma_{j}^{(C)}$
Non homogène	L'effet de la covariable sur les paramètres de seuil des items varie en magnitude, mais conserve la même direction
Taille du DIF	
Faible	
DIF homogène	$\forall\ p,\ \gamma_{jp}^{(C)} = \gamma_{j}^{(C)} = 0{,}3$
DIF non homogène	$\gamma_{j1}^{(C)} = 0{,}1$; $\gamma_{j2}^{(C)} = 0{,}3$; $\gamma_{j3}^{(C)} = 0{,}5$
Moyenne	
DIF homogène	$\forall\ p,\ \gamma_{jp}^{(C)} = \gamma_{j}^{(C)} = 0{,}5$
DIF non homogène	$\gamma_{j1}^{(C)} = 0{,}1$; $\gamma_{j2}^{(C)} = 0{,}5$; $\gamma_{j3}^{(C)} = 0{,}9$

Notes : $C = C_1$ *ou* C_2

5.5.3 Critères d'évaluation des performances

Nous avons utilisé plusieurs critères afin d'évaluer les performances des trois procédures en termes de détection du DIF, et ainsi déterminer quelle stratégie est la plus appropriée pour déterminer l'origine du DIF en présence de deux covariables binaires potentiellement corrélées.

Scénarios sans DIF

Pour chacun des huit scénarios sans DIF, nous avons utilisé le taux de détection à tort du DIF. Ce taux a été calculé pour chaque scénario comme la proportion de jeux de données où du DIF a été détecté à tort sur au moins une paire item-covariable (une fois la procédure étudiée terminée). Pour chacune des procédures, ce taux est espéré aussi bas que possible, sans pour autant qu'un seuil n'ait été prédéfini.

De plus, comme les algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD impliquent un test du rapport de vraisemblance effectué avec un $\alpha = 5\%$, nous avons également calculé la proportion de jeux de données ayant un test du rapport de vraisemblance significatif pour la confronter à la valeur théorique de 5%. Pour ces deux algorithmes, une différence entre la proportion de jeux de données ayant un test du rapport de vraisemblance significatif et le taux de détection à tort du DIF indique que pour certains jeux de données, la présence globale de DIF a été initialement suspectée à la suite du test du rapport de vraisemblance, mais que cette conclusion n'a finalement pas été retenue à la fin de la procédure.

Scénarios avec DIF

Pour les scénarios où du DIF a été simulé, un ensemble de critères a été utilisé pour évaluer les capacités des différentes procédures à retrouver ce qui avait été simulé. Ces critères sont présentés ci-dessous par niveau d'exigence croissant. Ils sont également représentés graphiquement dans la figure 5.8.

Le critère le plus flexible indique si la procédure a au moins permis de retrouver les paires item-covariable pour lesquelles du DIF a été simulé.

Le critère flexible indique si la procédure a permis de retrouver exactement les paires item-covariable pour lesquelles du DIF avait été simulé. Ce critère reprend les caractéristiques du critère le plus flexible, mais ajoute une contrainte : la procédure doit uniquement identifier les bonnes paires. Cela signifie qu'elle ne doit pas s'être trompée en identifiant également d'autres paires n'étant pas affectées par du DIF.

Le critère parfait indique si la procédure a permis de détecter exactement ce qui a été simulé (c.-à-d. que du DIF n'a été détecté que sur les bonnes paires item-covariable et que la forme du DIF été correctement déterminée).

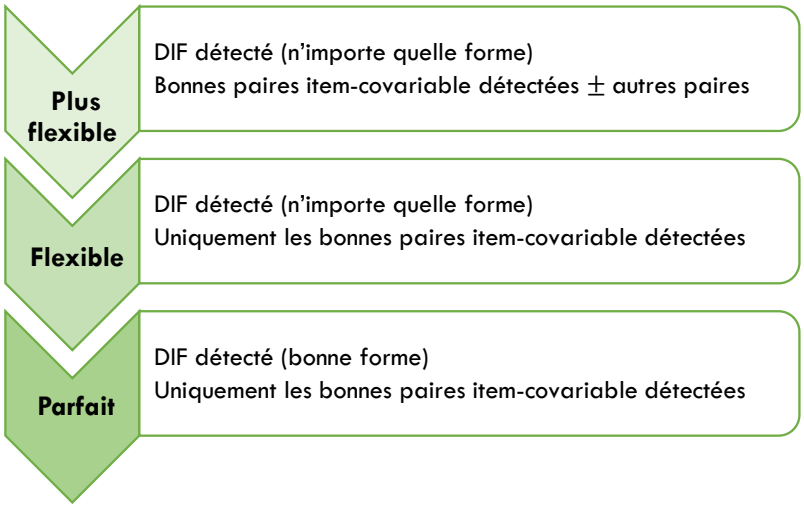


FIGURE 5.8 – Critères d'évaluation des performances pour la détection du DIF

Les performances des trois méthodes (ROSALI-DIF FORWARD, ROSALI-DIF BACKWARD et PCM-Lasso) ont été évaluées pour chaque scénario en estimant la proportion de jeux de données satisfaisant ces trois critères une fois la détection du DIF achevée. Ces proportions correspondent ainsi à des taux de détection à raison du DIF pour différents niveaux d'exigence.

Les différences entre les taux de détection à raison du DIF ont aussi été étudiées, car elles apportent une information complémentaire. En effet, pour une méthode donnée, la proportion de jeux de données satisfaisant le critère le plus flexible, mais ne satisfaisant pas le critère flexible indique dans quelle mesure la procédure a identifié à tort d'autres paires item-covariable en plus de celles pour lesquelles du DIF avait été simulé. De façon similaire, la proportion de jeux de données satisfaisant le critère flexible, mais pas le critère parfait indique dans quelle proportion la procédure a détecté les bonnes paires item-covariable (et seulement celles-ci) mais n'a pas correctement identifié la forme de DIF impliquée.

Nous avons ensuite comparé les estimations des paramètres DIF $\gamma_{jp}^{(C)}$ aux vraies valeurs simulées en utilisant des *box plots*. Enfin, nous avons évalué le biais dans l'estimation des effets des covariables sur le niveau moyen de la variable latente (notés β_1 et β_2 dans l'équation (5.4)) afin de déterminer si les trois méthodes permettent une estimation non biaisée de ces paramètres. En plus du biais, nous avons également estimé les erreurs standards empirique et asymptotique associées à l'estimation de $\beta_{k\ (k=1,2)}$. Ces mesures de performance sont définies dans le tableau 5.4 et illustrées graphiquement dans la figure 5.9.

TABLEAU 5.4 – Définition du biais et des erreurs standards empirique et asymptotique lors de l'estimation d'un paramètre β
Source : Adapté de Morris, Statistics in Medicine, 2019

Mesures de performance	**Définition**	**Estimation**
Biais	$E[\hat{\beta}] - \beta$	$\frac{1}{n_{\text{sim}}} \sum_{r=1}^{n_{\text{sim}}} \hat{\beta}^{(r)} - \beta$
Erreur standard empirique (eSE)	$\sqrt{\text{Var}(\hat{\beta})}$	$\sqrt{\frac{1}{n_{\text{sim}} - 1} \sum_{i=1}^{n_{\text{sim}}} \left(\hat{\beta}^{(r)} - \bar{\beta}\right)^2}$
Erreur standard asymptotique (aSE)	$\sqrt{E[\widehat{\text{Var}}(\hat{\beta})]}$	$\sqrt{\frac{1}{n_{\text{sim}}} \sum_{r=1}^{n_{\text{sim}}} \widehat{\text{Var}}\left(\hat{\beta}\right)^{(r)}}$

Notes :

β *: Vraie valeur du paramètre*

n_{sim} *: Nombre de jeux de données pour lesquels la méthode étudiée a convergé*

r *: Numéro du jeu de données (ou de la réplication)*

$\hat{\beta}$ *: L'estimateur de* β

$\hat{\beta}^{(r)}$ *: L'estimation de* β *pour le* $r^{ième}$ *jeu de données*

$\bar{\beta}$ *: La moyenne des* $\hat{\beta}^{(r)}$ *(les estimations* β*) sur l'ensemble des jeux de données*

$\text{Var}(\hat{\beta})$ *: La "vraie" variance de* $\hat{\beta}$

$\widehat{\text{Var}}\left(\hat{\beta}\right)^{(r)}$ *: L'estimation de la variance de* $\hat{\beta}$ *par le modèle pour le* $r^{ième}$ *jeu de données*

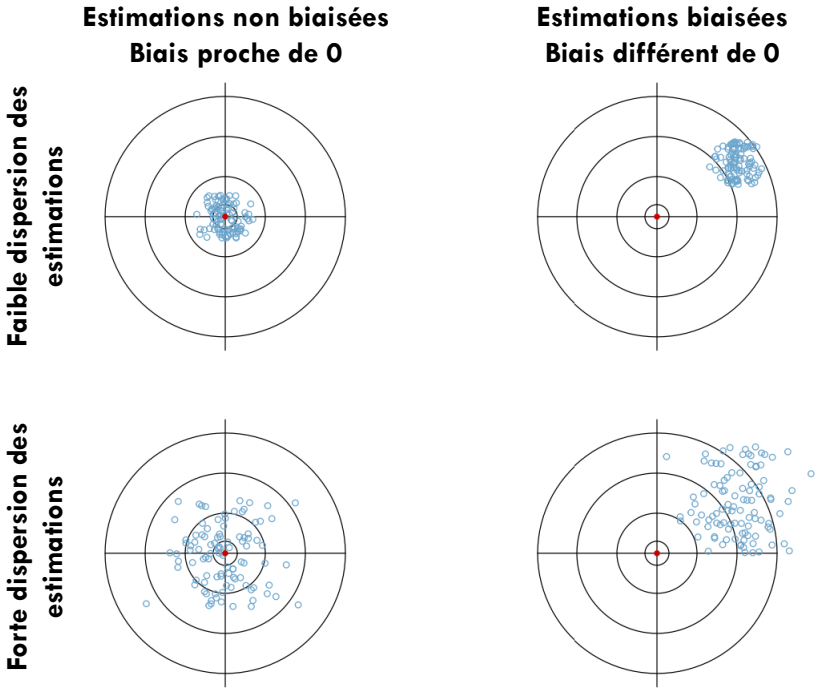


FIGURE 5.9 – Exemple de cas de figure lors de l'estimation d'un paramètre

Notes : La vraie valeur du paramètre est symbolisée par les points rouges et les estimations par les cercles bleus. Lorsque les estimations sont centrées autour de la vraie valeur du paramètre, on dit qu'elles sont non biaisées (biais à 0, colonne de gauche). Dans les autres cas (colonne de droite), les estimations ne sont pas centrées autour de la vraie valeur du paramètre : elles sont biaisées (biais ≠ 0). Lorsque les estimations sont peu dispersées (ligne du haut), l'erreur standard empirique est faible. Au contraire, lorsque les estimations sont plus dispersées (ligne du bas), l'erreur standard empirique est plus élevée.

5.5.4 Outils logiciels

Ces travaux ont été réalisés à l'aide des logiciels Stata version 16 (StataCorp, College station, TX) et R version 4.1.0. La simulation des données a été réalisée grâce au module Stata *simirt* qui permet de générer des données PRO selon différents modèles de la famille de Rasch ou de la famille de Lord (pour items dichotomiques et polytomiques) [163]. Les scripts qui permettent la détection du DIF grâce aux algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD ont été réalisés à partir du module Stata *gsem*. Pour la détection du DIF avec la méthode PCM-Lasso, nous avons eu recours au package GPCM-Lasso version 0.1-5 développé sur R par Schauberger et Mair [19]. L'ensemble des données simulées et des scripts sont disponibles en accès libre sur la plate-forme de science ouverte OSF (lien cliquable).

5.6 Résultats

5.6.1 Détection à tort - *Détection du DIF alors qu'il n'a pas été simulé*

Le tableau 5.5 présente les taux de détection à tort du DIF pour les trois méthodes évaluées (ROSALI-DIF FORWARD, ROSALI-DIF BACKWARD, PCM-Lasso), en fonction des caractéristiques des scénarios suivantes : la taille de l'échantillon N, le nombre d'items J et la présence ou l'absence de corrélation entre les deux covariables C_1 et C_2. Ce tableau indique également la proportion de jeux de données avec un test du rapport de vraisemblance significatif pour les algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD. Les trois procédures se sont terminées normalement pour l'ensemble des jeux de données.

Pour les algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD, la proportion de jeux de données où du DIF a été détecté à tort sur au moins une paire item-covariable à la fin de l'étape 6 est restée faible pour tous les scénarios (oscillant entre 3% et 6 %). Ni la taille de l'échantillon N, ni le nombre d'items J n'ont semblé impacter ces résultats. Néanmoins, les taux de détection du DIF à tort étaient généralement légèrement inférieurs lorsque les deux covariables C_1 et C_2 étaient corrélées (par rapport à quand elles ne l'étaient pas). Pour chaque scénario, le

taux de détection du DIF à tort était systématiquement inférieur à la proportion de jeux de données ayant un test du rapport de vraisemblance significatif. Cela signifie que pour certains jeux de données (de 1 à 3%), la présence globale de DIF a été initialement suspectée suite au test du rapport de vraisemblance, mais n'a finalement pas été retenue à la fin de l'algorithme. On peut enfin remarquer que les proportions de jeux de données ayant un test du rapport de vraisemblance significatif sont proches de 5% (le risque de première espèce utilisé pour ces tests).

Pour la procédure PCM-Lasso, du DIF a été détecté à tort sur au moins une paire item-covariable à la fin de la procédure pour près de la moitié des jeux de données, quel que soit le scénario considéré. Aucun des paramètres de simulation explorés n'a semblé impacter ces résultats. Malgré ces taux très élevés de détection à tort du DIF, nous avons choisi de continuer à explorer ses performances (pour la détection à raison du DIF) afin d'avoir un éventail complet de ses avantages et de ses limites.

TABLEAU 5.5 – Taux de détection à tort du DIF (%DIF détecté) et proportion de jeux de données avec un test du rapport de vraisemblance significatif (%LRT SIG) pour les scénarios sans DIF simulé.

			ROSALI-DIF FORWARD		**ROSALI-DIF BACKWARD**		**PCM-Lasso**
N	J	Corr	**%LRT SIG**	**%DIF Détecté**	**%LRT SIG**	**%DIF Détecté**	**%DIF Détecté**
400	4	Non	5%	4%	5%	4%	50%
400	4	Oui	6%	3%	6%	3%	50%
400	7	Non	7%	6%	7%	6%	44%
400	7	Oui	6%	3%	6%	4%	54%
800	4	Non	6%	4%	6%	4%	46%
800	4	Oui	5%	3%	5%	4%	47%
800	7	Non	6%	5%	6%	5%	48%
800	7	Oui	4%	3%	4%	3%	46%

Notes :

%LRT SIG : *Proportion de jeux de données avec un test du rapport de vraisemblance significatif*

%DIF Détecté : *Proportion de jeux de données où du DIF a été détecté à tort sur au moins une paire item-covariable à la fin de la procédure*

N : *Taille de l'échantillon*

J : *Nombre d'items*

Corr : *Corrélation entre C_1 et C_2*

5.6.2 Détection à raison - *Détection du DIF quand il a été simulé*

Les tableaux 5.6, 5.7 et 5.8 présentent les proportions de jeux de données satisfaisant les critères de détection "le plus flexible", "flexible" et "parfait" à la fin de chacune des trois procédures (ROSALI-DIF FORWARD, ROSALI-DIF BACKWARD et PCM-Lasso) pour les configurations 1, 2 et 3, respectivement. Ces trois procédures se sont terminées normalement sur l'ensemble des jeux de données.

Les trois procédures ont présenté des performances variables, qui dépendaient des valeurs des paramètres de simulation. Tout d'abord, parmi les scénarios où la taille de DIF était faible et la taille de l'échantillon était de 400, aucune des trois méthodes n'a réellement réussi à retrouver au moins partiellement le DIF simulé. En effet, parmi ces scénarios, les taux de détection à raison du DIF ne dépassaient pas 10%. Les résultats concernant ces scénarios ne seront donc pas développés davantage. Les paragraphes suivants se concentrent uniquement sur les résultats observés lorsque la taille de l'échantillon N était égale à 800 ou lorsque la taille du DIF était moyenne (avec soit N = 400 ou 800).

Configurations n° 1 et 2 (C_1 et C_2 non corrélées induisant toutes deux du DIF)

Quelle que soit la procédure considérée, les taux de détection avec le critère le plus flexible étaient faibles lorsque la taille du DIF était faible et la taille d'échantillon valait 800 (taux compris entre 16% et 34%). Lorsque la taille du DIF était moyenne, les taux de détection avec le critère le plus flexible augmentaient. En effet, ils étaient modérés lorsque N = 400 (entre 35% et 56%) et élevés lorsque N = 800 (entre 76% à 96%). Pour rappel, ces taux indiquent dans quelle mesure les différentes procédures ont été capables d'identifier du DIF sur les paires item-covariable pour lesquelles du DIF avait effectivement été simulé (sans exiger que les procédures ne détecte pas en plus du DIF sur d'autres paires). Pour le critère le plus flexible, les meilleures performances ont été observées pour l'algorithme ROSALI-DIF FORWARD (taux de détection à raison allant de 20% à 96%, moyenne : 56%), mais l'algorithme ROSALI-DIF BACKWARD a démontré des performances assez similaires (les taux de détection à raison ne différaient pas de plus de 5%). Les per-

formances de la procédure PCM-Lasso étaient généralement légèrement inférieures (taux de détection avec le critère le plus flexible oscillant entre 16% et 90%, moyenne : 50%). On peut remarquer que les trois méthodes présentaient des taux de détection à raison généralement plus élevés lorsque le DIF était non homogène que lorsqu'il était homogène, tous les autres paramètres de simulation égaux par ailleurs, avec une différence maximale de +17% (différence moyenne de +8%).

En augmentant le niveau d'exigence et en s'intéressant au critère flexible, les taux de détection pour les algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD étaient : faibles lorsque la taille du DIF était faible et N = 800 (de 13% à 32%), modérés lorsque la taille du DIF était moyenne et N = 400 (de 30% à 49%) et élevés lorsque la taille du DIF était moyenne et N = 800 (de 65% à 81% et de 78% à 87% pour ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD, respectivement). En ce qui concerne la procédure PCM-Lasso, nous avons observé de mauvais taux de détection avec le critère flexible pour tous les scénarios, exceptés ceux avec une taille d'échantillon N = 800 et une taille DIF moyenne (les taux variaient alors entre 37% à 54%).

Par conséquent, en se basant sur le critère flexible, les méthodes les plus performantes sont les algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD. On peut par ailleurs remarquer que l'algorithme ROSALI-DIF BACKWARD est plus performant que l'algorithme ROSALI-DIF FORWARD lorsque la taille du DIF est moyenne et que N = 800 (les taux de détection des deux méthodes avec le critère flexible différaient dans ce cas de +5% à +16%), tandis que leurs performances étaient similaires parmi les autres scénarios. Enfin, les trois méthodes ont démontré des taux de détection avec le critère flexible généralement plus élevés lorsque le DIF était non homogène que lorsqu'il était homogène (tous les autres paramètres de simulation égaux par ailleurs), avec une différence maximale de +17% (différence moyenne de +7%).

Pour chacune des procédures, la proportion de jeux de données satisfaisant le critère flexible était systématiquement inférieure à la proportion de jeux de données satisfaisant le critère le plus flexible. Cela signifie qu'en plus des paires item-covariable effectivement affectées par du

DIF, toutes les procédures ont détecté à tort d'autres paires (sur lesquelles aucun DIF n'avait été simulé). Cet écart était d'autant plus grand que la proportion de jeux de données satisfaisant le critère le plus flexible était grand. Pour les algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD, l'écart entre les deux taux de détection augmentait également avec le nombre d'items J. Par ailleurs, ces écarts étaient toujours les plus faibles pour l'algorithme ROSALI-DIF BACKWARD et les plus grands pour PCM-Lasso.

Parmi les scénarios avec du DIF homogène, les algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD ont tous deux présenté des taux de détection avec le critère parfait proches de ceux avec le critère flexible (par exemple, ces taux différaient de 4% à 10% pour les scénarios avec du DIF moyen). Cela indique que parmi les jeux de données satisfaisant le critère flexible, les deux algorithmes ont correctement déterminé la forme de DIF impliquée lorsque celle-ci était homogène. Pour la procédure PCM-Lasso, les taux de détection avec le critère parfait ne dépassaient pas 2%. Cela signifie que la procédure n'a pas réussi à identifier la bonne forme de DIF lorsque le DIF simulé était homogène. Par conséquent, sur la base du critère parfait et parmi les scénarios avec du DIF homogène, les algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD ont démontré des performances supérieures à celles de la procédure PCM-Lasso. En revanche, parmi les scénarios avec du DIF non homogène, les taux de détection avec le critère parfait pour les algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD étaient nettement inférieurs à ceux avec le critère flexible (par exemple, les écarts allaient de 30% à 56% parmi les scénarios avec du DIF moyen). Cela signifie que parmi les jeux de données satisfaisant le critère flexible, les deux algorithmes n'ont pas réussi à correctement identifier la forme du DIF simulée (c.-à-d. du DIF non homogène). L'approche PCM-Lasso a, au contraire, présenté des taux de détection avec le critère parfait très proches des taux de détection avec le critère flexible, c'est-à-dire des taux modérés allant de 49% à 53% lorsque le DIF était moyen et $N = 800$, et des taux faibles allant de 12% à 26% dans les autres cas. Cela signifie qu'une fois que l'approche PCM-Lasso a correctement identifié les paires item-covariable affectées par du DIF, elle a également identifié la bonne forme de DIF. Par conséquent, lorsque le DIF n'était pas homogène, l'approche

PCM-Lasso présentait des taux de détection avec le critère parfait plus élevés que les deux autres algorithmes. Ainsi, sur la base du critère parfait, l'approche PCM-Lasso a surpassé ROSALI-DIF FORWARD et ROSALI-DIF BACKAWARD lorsque le DIF simulé était non homogène. On peut enfin noter que l'algorithme ROSALI-DIF BACKWARD a présenté des taux de détection avec le critère parfait plus élevés que ROSALI-DIF FORWARD parmi les scénarios avec du DIF de taille moyenne et une taille d'échantillon N = 800. Les deux algorithmes ont montré des performances similaires dans les autres cas. Cet effet avait déjà été observé avec le critère flexible.

Configuration n° 3 (C_1 et C_2 corrélées, seule C_1 induit du DIF sur deux items)

Lorsque le DIF était simulé selon la configuration n°3, les performances des trois méthodes étaient presque toujours moins bonnes que pour les deux premières configurations, quel que soit le critère considéré. Par exemple, en s'intéressant aux critères de détection le plus flexible et flexible, les performances des trois méthodes étaient globalement faibles lorsque la taille du DIF était faible ou lorsque la taille de l'échantillon était égale à 400. Pour les algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD, ces taux étaient meilleurs parmi les scénarios avec N = 800 et un DIF de taille moyenne (allant de 69% et 86% avec le critère le plus flexible et de 61% à 80% pour le critère flexible). Les performances étaient moins bonnes pour l'approche PCM-Lasso dans les mêmes conditions : les taux de détection avec le critère le plus flexible oscillaient entre 19% et 78% et les taux de détection avec le critère flexible allaient de 3% à 34%.

Des effets similaires à ceux mis en évidence dans les configurations n° 1 et 2 ont été observés, à savoir :

- Les taux de détection avec le critère le plus flexible et le critère flexible étaient plus élevés lorsque la forme du DIF simulé était non homogène.
- Les taux de détection avec le critère flexible étaient inférieurs aux taux de détection avec le critère le plus flexible pour toutes les procédures, ce qui indique qu'en plus des bonnes paires item-covariable, toutes les procédures ont détecté du DIF à tort sur d'autres paires. L'écart entre ces taux était d'autant plus grand que le taux de détection avec le critère

le plus flexible était élevé. En outre, pour les algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD, l'écart augmentait également lorsque le nombre d'items J passait de 4 à 7. Ces écarts étaient toujours les plus faibles pour l'algorithme ROSALI-DIF BACKWARD et les plus importants pour l'approche PCM-Lasso.

– La méthode PCM-Lasso a presque toujours mis en évidence du DIF non homogène (quelle que soit la forme de DIF simulée). Par conséquent, les taux de détection avec le critère parfait ne dépassaient pas 1% parmi les scénarios où le DIF avait été simulé homogène. En revanche, au sein des scénarios avec du DIF non homogène, les taux de détection avec le critère parfait étaient presque équivalents aux taux de détection avec le critère flexible.

– Les deux algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD ont de nouveau bien identifié le DIF homogène, mais n'ont pas réussi à identifier la bonne forme de DIF quand le DIF était non homogène (les écarts entre les taux de détection avec les critères flexible et parfait étant grands).

TABLEAU 5.6 – Taux de détection à raison du DIF en fonction des différents niveaux d'exigence parmi les scénarios où du DIF a été simulé selon la configuration 1 (les deux covariables ne sont pas corrélées et elles induisent toutes deux du DIF sur un item différent). Les résultats sont donnés en fonction des paramètres de simulation suivants : forme du DIF (H : Homogène, NH : non homogène), taille du DIF, taille de l'échantillon N et nombre d'items J.

				ROSALI-DIF FORWARD				**ROSALI-DIF BACKWARD**				**PCM-Lasso**		
DIF : Forme	DIF : Taille	N	J	**%LRT SIG**	**%Plus flexible**	**% Flexible**	**% Parfait**	**%LRT SIG**	**%Plus flexible**	**% Flexible**	**% Parfait**	**%Plus flexible**	**% Flexible**	**% Parfait**
H	Faible	400	4	31%	4%	3%	2%	31%	3%	3%	1%	5%	3%	0%
H	Faible	400	7	27%	5%	4%	3%	27%	5%	4%	3%	6%	3%	0%
H	Faible	800	4	64%	20%	17%	14%	64%	16%	15%	12%	16%	9%	1%
H	Faible	800	7	63%	26%	18%	14%	63%	24%	21%	17%	18%	9%	0%
H	Moyenne	400	4	85%	39%	34%	29%	85%	36%	33%	29%	35%	17%	0%
H	Moyenne	400	7	79%	44%	30%	25%	79%	39%	35%	30%	39%	16%	0%
H	Moyenne	800	4	99%	90%	73%	67%	99%	88%	81%	73%	80%	42%	1%
H	Moyenne	800	7	99%	91%	66%	59%	99%	91%	81%	73%	83%	46%	2%
NH	Faible	400	4	33%	5%	4%	1%	33%	4%	4%	0%	8%	6%	5%
NH	Faible	400	7	31%	4%	2%	0%	31%	4%	3%	0%	6%	3%	3%
NH	Faible	800	4	68%	29%	25%	3%	68%	25%	23%	3%	27%	18%	17%
NH	Faible	800	7	62%	27%	18%	2%	62%	24%	21%	3%	23%	14%	12%
NH	Moyenne	400	4	91%	56%	49%	6%	91%	51%	49%	6%	49%	26%	25%
NH	Moyenne	400	7	79%	51%	35%	5%	79%	48%	41%	6%	49%	24%	23%
NH	Moyenne	800	4	100%	96%	77%	31%	100%	94%	87%	35%	88%	49%	49%
NH	Moyenne	800	7	100%	96%	65%	24%	100%	95%	80%	30%	90%	52%	51%

Notes :

%LRT SIG : *Proportion de jeux de données avec un test du rapport de vraisemblance significatif*

%Plus flexible : *Proportion de jeux de données où la procédure a détecté du DIF sur au moins les bonnes paires item-covariable (parmi d'autres)*

%Flexible : *Proportion de jeux de données où la procédure a détecté du DIF uniquement sur les bonnes paires item-covariable*

%Parfait : *Proportion de jeux de données où la procédure a retrouvé exactement le DIF simulé (bonnes paires uniquement et bonne forme de DIF)*

TABLEAU 5.7 – Taux de détection à raison du DIF en fonction des différents niveaux d'exigence parmi les scénarios où du DIF a été simulé selon la configuration 2 (les deux covariables ne sont pas corrélées et elles induisent toutes deux du DIF sur le même item). Les résultats sont donnés en fonction des paramètres de simulation suivants : forme du DIF (H : Homogène, NH : non homogène), taille du DIF, taille de l'échantillon N et nombre d'items J.

				ROSALI-DIF FORWARD				**ROSALI-DIF BACKWARD**				**PCM-Lasso**		
DIF : Forme	DIF : Taille	N	J	**%LRT SIG**	**%Plus flexible**	**% Flexible**	**% Parfait**	**%LRT SIG**	**%Plus flexible**	**% Flexible**	**% Parfait**	**%Plus flexible**	**% Flexible**	**% Parfait**
H	Faible	400	4	29%	3%	1%	1%	29%	2%	1%	1%	5%	3%	0%
H	Faible	400	7	26%	2%	2%	1%	26%	2%	2%	1%	6%	2%	0%
H	Faible	800	4	68%	23%	20%	18%	68%	20%	19%	16%	17%	9%	0%
H	Faible	800	7	58%	21%	13%	10%	58%	19%	16%	12%	17%	8%	0%
H	Moyenne	400	4	83%	41%	35%	30%	83%	38%	36%	31%	35%	15%	0%
H	Moyenne	400	7	82%	41%	30%	25%	82%	39%	33%	28%	43%	15%	1%
H	Moyenne	800	4	99%	91%	77%	68%	99%	89%	84%	74%	76%	37%	1%
H	Moyenne	800	7	99%	92%	65%	61%	99%	92%	78%	73%	85%	46%	2%
NH	Faible	400	4	34%	6%	6%	1%	34%	4%	4%	1%	10%	6%	6%
NH	Faible	400	7	32%	7%	5%	1%	32%	6%	6%	1%	9%	5%	4%
NH	Faible	800	4	74%	34%	32%	2%	74%	30%	29%	2%	26%	16%	15%
NH	Faible	800	7	67%	32%	22%	2%	67%	29%	25%	3%	27%	15%	13%
NH	Moyenne	400	4	86%	50%	42%	5%	86%	46%	43%	6%	47%	27%	26%
NH	Moyenne	400	7	79%	51%	39%	6%	79%	48%	42%	7%	46%	22%	22%
NH	Moyenne	800	4	100%	95%	81%	26%	100%	93%	86%	30%	86%	54%	53%
NH	Moyenne	800	7	99%	96%	65%	22%	99%	95%	81%	28%	89%	51%	50%

Notes :

%LRT SIG : *Proportion de jeux de données avec un test du rapport de vraisemblance significatif*

%Plus flexible : *Proportion de jeux de données où la procédure a détecté du DIF sur au moins les bonnes paires item-covariable (parmi d'autres)*

%Flexible : *Proportion de jeux de données où la procédure a détecté du DIF uniquement sur les bonnes paires item-covariable*

%Parfait : *Proportion de jeux de données où la procédure a retrouvé exactement le DIF simulé (bonnes paires uniquement et bonne forme de DIF)*

TABLEAU 5.8 – Taux de détection à raison du DIF en fonction des différents niveaux d'exigence parmi les scénarios où du DIF a été simulé selon la configuration 3 (les deux covariables sont corrélées, une seule induit du DIF sur deux items). Les résultats sont donnés en fonction des paramètres de simulation suivants : forme du DIF (H : Homogène, NH : non homogène), taille du DIF, taille de l'échantillon N et nombre d'items J.

				ROSALI-DIF FORWARD				**ROSALI-DIF BACKWARD**				**PCM-Lasso**		
DIF : Forme	DIF : Taille	N	J	**%LRT SIG**	**%Plus flexible**	**% Flexible**	**% Parfait**	**%LRT SIG**	**%Plus flexible**	**% Flexible**	**% Parfait**	**%Plus flexible**	**% Flexible**	**% Parfait**
H	Faible	400	4	21%	2%	2%	1%	21%	3%	2%	1%	2%	0%	0%
H	Faible	400	7	23%	2%	2%	2%	23%	2%	1%	1%	1%	1%	0%
H	Faible	800	4	44%	7%	7%	6%	44%	8%	7%	6%	2%	0%	0%
H	Faible	800	7	44%	13%	10%	8%	44%	13%	11%	9%	9%	3%	0%
H	Moyenne	400	4	58%	19%	18%	13%	58%	20%	20%	16%	7%	2%	0%
H	Moyenne	400	7	67%	26%	22%	18%	67%	26%	24%	19%	20%	6%	0%
H	Moyenne	800	4	91%	69%	61%	54%	91%	71%	66%	58%	19%	3%	0%
H	Moyenne	800	7	96%	78%	62%	56%	96%	77%	70%	62%	60%	21%	1%
NH	Faible	400	4	24%	3%	2%	0%	24%	3%	2%	0%	3%	1%	1%
NH	Faible	400	7	28%	2%	1%	0%	28%	2%	1%	0%	4%	1%	1%
NH	Faible	800	4	48%	13%	11%	1%	48%	13%	13%	1%	8%	4%	4%
NH	Faible	800	7	52%	17%	13%	2%	52%	17%	15%	2%	15%	6%	6%
NH	Moyenne	400	4	73%	31%	28%	5%	73%	33%	30%	5%	22%	8%	8%
NH	Moyenne	400	7	75%	38%	31%	6%	75%	36%	32%	6%	34%	16%	16%
NH	Moyenne	800	4	98%	86%	73%	25%	98%	85%	80%	31%	58%	19%	19%
NH	Moyenne	800	7	98%	85%	66%	26%	98%	85%	74%	29%	78%	34%	33%

Notes :

%LRT SIG : *Proportion de jeux de données avec un test du rapport de vraisemblance significatif*

%Plus flexible : *Proportion de jeux de données où la procédure a détecté du DIF sur au moins les bonnes paires item-covariable (parmi d'autres)*

%Flexible : *Proportion de jeux de données où la procédure a détecté du DIF uniquement sur les bonnes paires item-covariable*

%Parfait : *Proportion de jeux de données où la procédure a retrouvé exactement le DIF simulé (bonnes paires uniquement et bonne forme de DIF)*

5.6.3 Estimation des paramètres de DIF

Les paramètres de DIF étaient systématiquement sous-estimés par la méthode PCM-Lasso alors qu'ils étaient globalement bien retrouvés par les algorithmes ROSALI-DIF (les estimations étant plus proches des vraies valeurs simulées). Ces résultats sont illustrés par les figures 5.10 et 5.11 pour deux scénarios définis par les caractéristiques suivantes :

- Taille d'échantillon $N = 400$;
- Nombre d'items $J = 4$ items ;
- Le DIF est simulé selon la configuration n°1, c.-à-d. les covariables C_1 et C_2 sont indépendantes, la covariable C_1 induit du DIF sur l'item 2 et la covariable C_2 induit du DIF sur l'item 3 ;
- Le DIF induit est homogène (figure 5.10) ou non homogène (figure 5.11) et de taille moyenne.

Ces deux scénarios ont été sélectionnés parce qu'ils présentaient des résultats intermédiaires en termes de détection à raison du DIF.

Les figures 5.10 et 5.11 contiennent trois lignes. La première contient les *box plots* des estimations des paramètres de DIF $\gamma_{2p}^{(C_1)}$ $\gamma_{3p}^{(C_2)}$ sur l'ensemble des 500 jeux de données associés au scénario considéré. La deuxième ligne se restreint aux jeux de données satisfaisant le critère le plus flexible (du DIF a été détecté pour au moins les bonnes paires item-covariable) et la troisième ligne se concentre sur les jeux de données satisfaisant le critère parfait (seules les bonnes paires ont été détectées et la forme du DIF a été correctement identifiée).

Les résultats pour les autres scénarios sont disponibles depuis une application R Shiny disponible depuis ce lien (lien cliquable).

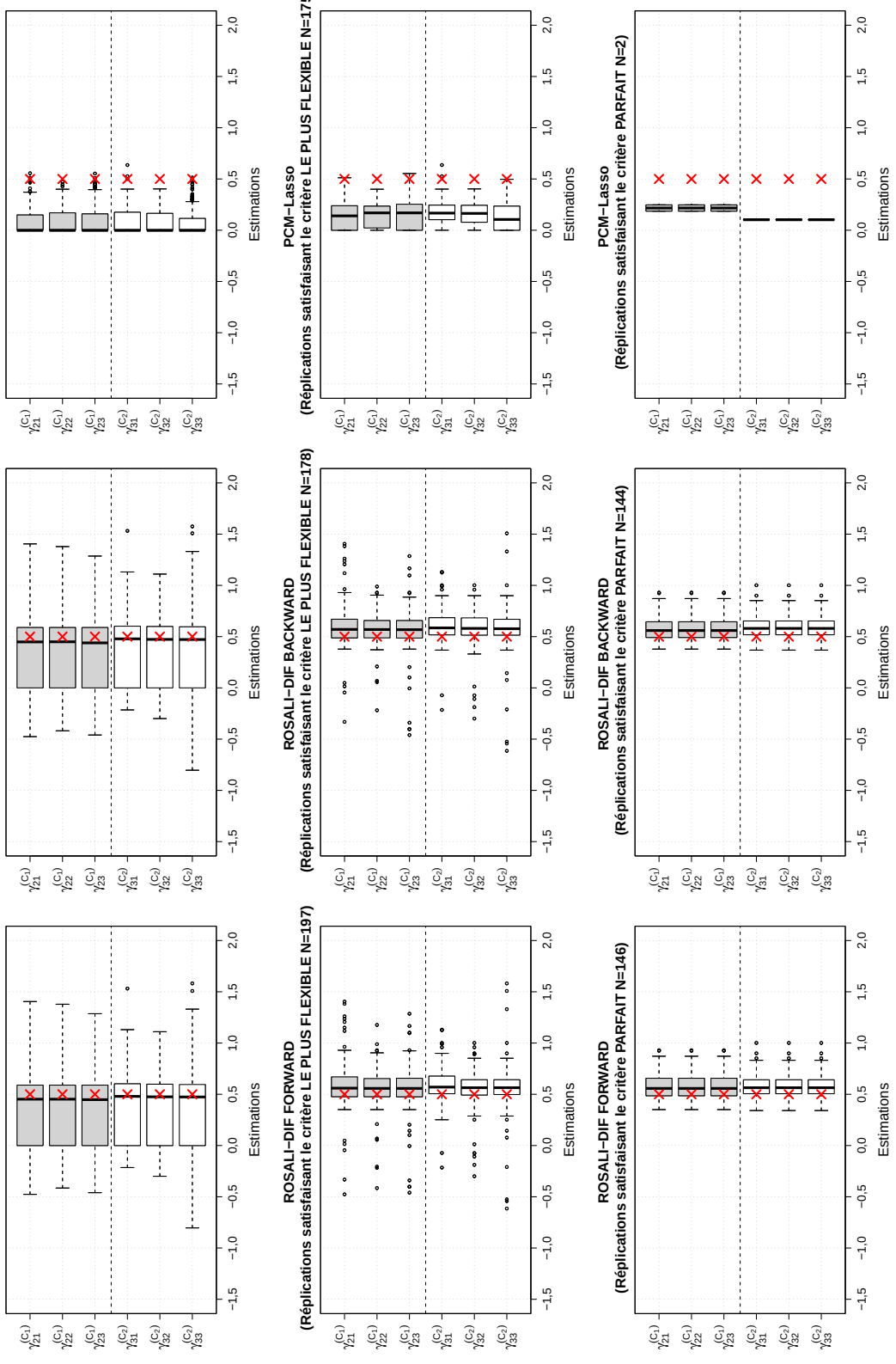


FIGURE 5.10 – Estimations des paramètres de DIF $\gamma_{jp}^{(C)}$ obtenues avec ROSALI-DIF FORWARD, ROSALI-DIF BACKWARD et PCM-Lasso pour le scénario où $N = 400$, $J = 4$ et C_1 et C_2 induisent respectivement du DIF homogène de taille moyenne sur les items 2 et 3 .
Notes : Les vraies valeurs simulées sont indiquées par les croix rouges

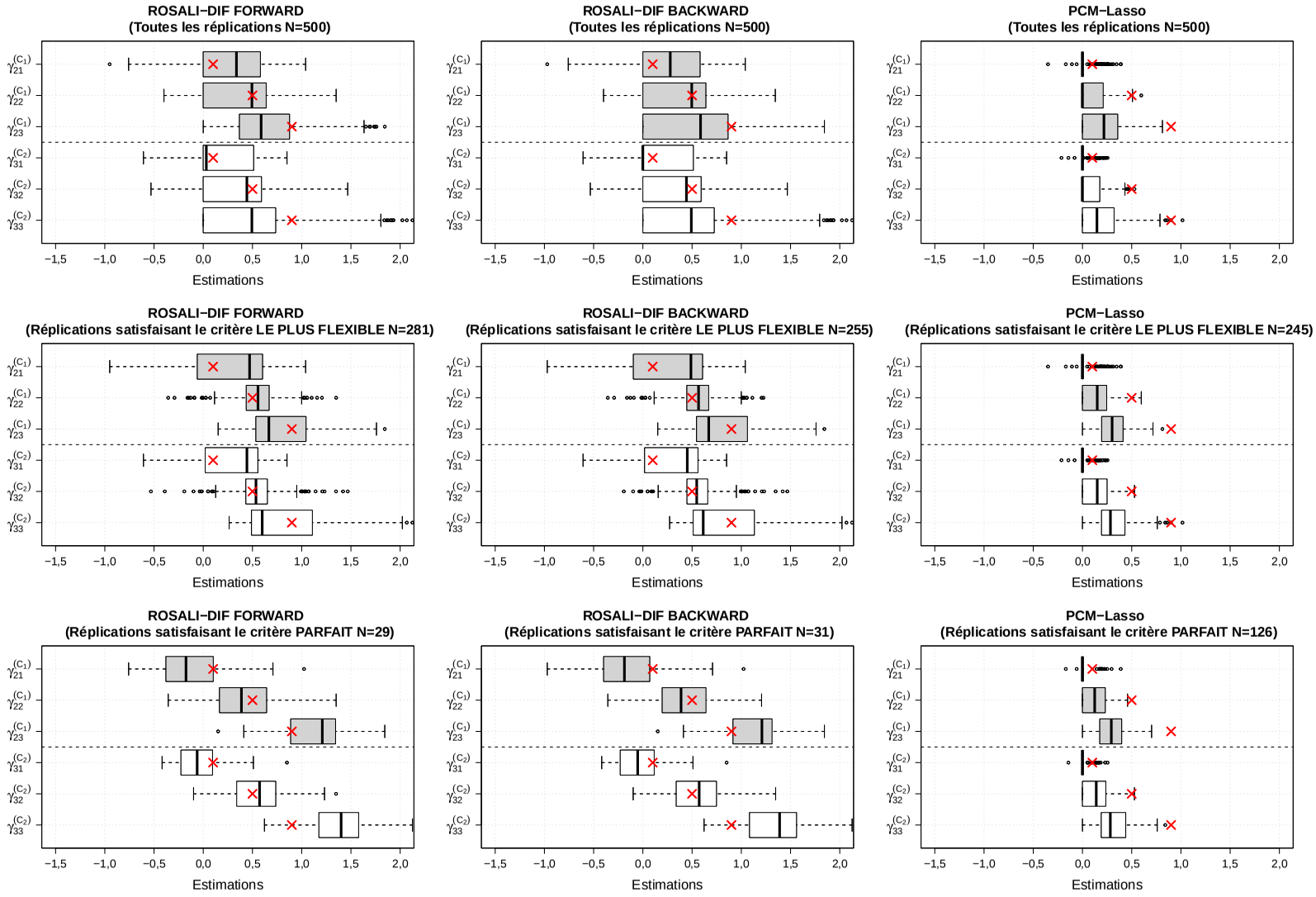


FIGURE 5.11 – Estimations des paramètres de DIF $\gamma_{jp}^{(C)}$ pour le scénario où $N = 400$, $J = 4$ et C_1 et C_2 induisent respectivement du DIF non homogène de taille moyenne sur les items 2 et 3 obtenues avec ROSALI-DIF FORWARD, ROSALI-DIF BACKWARD et PCM-Lasso.

Notes : Les vraies valeurs simulées sont indiquées par les croix rouges

5.6.4 Estimation de l'effet des covariables sur le niveau de la variable latente parmi les scénarios avec DIF

Biais

À la fin des trois procédures étudiées, les paramètres β_1 et β_2 (représentant l'effet des covariables C_1 et C_2 sur le niveau moyen de la variable latente) sont estimés en ajustant sur le DIF mis en évidence. Le biais dans l'estimation de β_1 et de β_2 est donné dans le tableau 5.9 pour les scénarios où le DIF a été simulé selon les configurations n°1 et 2 et dans le tableau 5.10 pour les scénarios où le DIF a été simulé selon la configuration n°3.

Pour les scénarios où les covariables C_1 et C_2 n'étaient pas corrélées et induisaient toutes deux du DIF (configurations n°1 et 2), le biais est resté faible pour toutes les méthodes : il n'a jamais dépassé 0,08 en valeur absolue.

Pour les scénarios où les covariables étaient corrélées et seule la covariable C_1 induisait du DIF (configuration n°3), les résultats étaient plus mitigés. En effet, le biais associé à l'estimation de β_1 (l'effet de la covariable C_1) est resté faible parmi les scénarios où J = 7, mais il était plus conséquent parmi les scénarios où J = 4, atteignant respectivement -0,19 et -0,18 pour les algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD, et -0,14 pour PCM-Lasso. L'estimation du paramètre β_2 (l'effet de la covariable C_2) n'était en revanche pas biaisée.

Sur l'ensemble des scénarios, on peut remarquer que plus les procédures détectaient bien le DIF simulé, plus le biais était faible.

TABLEAU 5.9 – Biais associé à l'estimation de β_1 et β_2 suite aux procédures de détection du DIF pour les scénarios où du DIF a été simulé selon les configurations n°1 et 2. Les résultats sont donnés en fonction des paramètres de simulation suivants : forme du DIF (H : Homogène, NH : non homogène), taille du DIF, taille de l'échantillon N et nombre d'items J.

				CONFIGURATION N°1						**CONFIGURATION N°2**					
				ROSALI-DIF FORWARD		**ROSALI-DIF BACKWARD**		**PCM-Lasso**		**ROSALI-DIF FORWARD**		**ROSALI-DIF BACKWARD**		**PCM-Lasso**	
DIF :	DIF :			**Biais**		**Biais**		**Biais**		**Biais**		**Biais**		**Biais**	
Forme	Taille	N	J	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
H	Faible	400	4	-0,07	-0,07	-0,07	-0,07	-0,04	-0,04	-0,06	-0,08	-0,06	-0,08	-0,03	-0,04
H	Faible	400	7	-0,04	-0,04	-0,04	-0,04	-0,02	-0,02	-0,04	-0,04	-0,04	-0,04	-0,02	-0,02
H	Faible	800	4	-0,05	-0,05	-0,05	-0,05	-0,04	-0,03	-0,04	-0,04	-0,04	-0,04	-0,03	-0,03
H	Faible	800	7	-0,02	-0,02	-0,02	-0,02	-0,02	-0,02	-0,02	-0,03	-0,02	-0,03	-0,02	-0,02
H	Moyenne	400	4	-0,05	-0,04	-0,05	-0,05	-0,05	-0,05	-0,04	-0,04	-0,04	-0,04	-0,04	-0,04
H	Moyenne	400	7	-0,03	-0,03	-0,03	-0,03	-0,03	-0,03	-0,02	-0,01	-0,02	-0,01	-0,02	-0,02
H	Moyenne	800	4	-0,02	-0,01	-0,02	-0,01	-0,04	-0,03	-0,00	-0,01	-0,00	-0,01	-0,03	-0,03
H	Moyenne	800	7	-0,00	-0,00	-0,00	-0,00	-0,02	-0,02	-0,00	0,00	-0,00	0,00	-0,02	-0,01
NH	Faible	400	4	-0,07	-0,07	-0,07	-0,07	-0,04	-0,04	-0,05	-0,06	-0,05	-0,06	-0,03	-0,03
NH	Faible	400	7	-0,03	-0,04	-0,03	-0,04	-0,02	-0,02	-0,03	-0,04	-0,03	-0,04	-0,01	-0,02
NH	Faible	800	4	-0,04	-0,04	-0,04	-0,04	-0,03	-0,03	-0,04	-0,04	-0,04	-0,04	-0,03	-0,03
NH	Faible	800	7	-0,01	-0,03	-0,01	-0,03	-0,01	-0,02	-0,02	-0,02	-0,02	-0,02	-0,01	-0,02
NH	Moyenne	400	4	-0,02	-0,02	-0,02	-0,02	-0,04	-0,04	-0,02	-0,02	-0,02	-0,02	-0,03	-0,03
NH	Moyenne	400	7	-0,01	-0,02	-0,01	-0,02	-0,02	-0,02	-0,02	-0,02	-0,02	-0,02	-0,02	-0,02
NH	Moyenne	800	4	0,00	-0,00	0,00	-0,00	-0,02	-0,03	0,00	-0,00	0,00	-0,01	-0,02	-0,03
NH	Moyenne	800	7	-0,00	-0,01	-0,00	-0,01	-0,02	-0,02	0,01	0,00	0,00	0,00	-0,01	-0,01

Notes :

β_1 et β_2 désignent l'effet respectif des covariables C_1 et C_2 sur le niveau moyen de la variable latente

Configuration n°1 : C_1 et C_2 ne sont pas corrélées, elles induisent toutes deux du DIF sur un item distinct

Configuration n°2 : C_1 et C_2 ne sont pas corrélées, elles induisent toutes deux du DIF sur le même item

TABLEAU 5.10 – Biais associé à l'estimation de β_1 et β_2 suite aux procédures de détection du DIF pour les scénarios où du DIF a été simulé selon la configuration n°3. Les résultats sont donnés en fonction des paramètres de simulation suivants : forme du DIF (H : Homogène, NH : non homogène), taille du DIF, taille de l'échantillon N et nombre d'items J.

				CONFIGURATION N°3					
				ROSALI-DIF FORWARD		**ROSALI-DIF BACKWARD**		**PCM-Lasso**	
DIF : Forme	DIF : Taille	N	J	**Biais** β_1	β_2	**Biais** β_1	β_2	**Biais** β_1	β_2
H	Faible	400	4	-0,14	-0,01	-0,14	-0,01	-0,08	-0,00
H	Faible	400	7	-0,08	-0,00	-0,08	-0,00	-0,04	-0,00
H	Faible	800	4	-0,13	-0,00	-0,12	-0,00	-0,08	-0,00
H	Faible	800	7	-0,06	-0,00	-0,06	-0,00	-0,04	-0,00
H	Moyenne	400	4	-0,19	0,00	-0,18	0,00	-0,14	0,00
H	Moyenne	400	7	-0,09	0,00	-0,09	0,00	-0,07	0,00
H	Moyenne	800	4	-0,05	0,00	-0,05	0,00	-0,13	0,00
H	Moyenne	800	7	-0,02	0,00	-0,02	0,00	-0,05	0,00
NH	Faible	400	4	-0,14	-0,00	-0,14	-0,00	-0,07	-0,00
NH	Faible	400	7	-0,06	-0,00	-0,06	-0,00	-0,03	-0,00
NH	Faible	800	4	-0,11	0,00	-0,11	0,00	-0,07	0,00
NH	Faible	800	7	-0,06	-0,00	-0,06	-0,00	-0,04	-0,00
NH	Moyenne	400	4	-0,14	-0,00	-0,13	-0,00	-0,11	-0,00
NH	Moyenne	400	7	-0,06	0,01	-0,06	0,01	-0,05	0,00
NH	Moyenne	800	4	-0,02	0,00	-0,02	0,00	-0,09	0,00
NH	Moyenne	800	7	-0,01	0,00	-0,01	0,00	-0,04	0,00

Notes :

β_1 et β_2 désignent l'effet respectif des covariables C_1 et C_2 sur le niveau moyen de la variable latente

Configuration n°3 : C_1 et C_2 sont corrélées, seule C_1 induit du DIF sur deux items

Les cellules grisées correspondent aux cas où le biais excédait 0,1 en valeur absolue

Erreurs standards empiriques et asymptotiques

Les erreurs standards empiriques se sont avérées systématiquement plus grandes pour les algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD que pour PCM-Lasso, avec un rapport multiplicatif proche de 2. Plus précisément, les erreurs standards empiriques associées aux deux premiers algorithmes variaient de 0,07 à 0,17 alors que celles de PCM-Lasso étaient comprises entre 0,04 et 0,08. Cela signifie que les estimations des paramètres β_1 et β_2 étaient plus dispersées pour les deux premières méthodes que pour la dernière. On peut néanmoins remarquer que les erreurs standards empiriques et asymptotiques obtenues avec les algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD étaient très proches (ne différant pas de plus de 0,02), n'indiquant pas de problème particulier dans l'estimation des erreurs standards de β_1 et β_2. Cette comparaison n'a pas pu être réalisée pour la méthode PCM-Lasso, les erreurs standards asymptotiques n'étant à notre connaissance pas fournies par le package GPCMLasso.

5.7 Discussion

Ces travaux visaient à étendre la première partie de l'algorithme ROSALI dédiée à la détection du DIF, afin de considérer deux covariables binaires au lieu d'une. Deux versions ont été proposées pour l'extension : ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD. La nouveauté qui caractérise ces extensions est l'étape 4 de *screening* qui vise à identifier les paires item-covariable candidates à la détection du DIF et les paires item-covariable qui seront considérées comme *anchor* (non affectées par du DIF). L'insertion de cette étape supplémentaire s'inspire de la procédure itérative proposée par Tay *et al.* lors de la recherche de DIF avec une covariable [89, 179]. D'après ces auteurs, tester la présence de DIF pour tous les items à partir d'un modèle complètement non-invariant présentait une bonne puissance, mais également un risque de première espèce élevé. Ils indiquaient donc que cette pratique pourrait être utile pour identifier les items *anchor* et les items candidats pour la détection du DIF lors d'une étape préliminaire [89, 179].

Les deux algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD ont été évalués par simulations, en parallèle de l'approche basée sur la pénalisation de la vraisemblance proposée par Schauberger et Mair [19] (PCM-Lasso) dans des conditions qui se voulaient représentatives du contexte de recherche en santé. L'objectif était de déterminer si ces méthodes étaient capables de retrouver le DIF lorsqu'il était simulé et n'inféraient pas de DIF à tort lorsque aucun fonctionnement différentiel n'était simulé.

Résultats principaux

D'après les taux de détection à tort, les deux algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD n'infèrent généralement pas de DIF à tort parmi les scénarios sans DIF. Ces bonnes performances s'expliquent probablement par le test du rapport de vraisemblance effectué à un risque de première espèce de 5%, l'étape de "screening", et l'étape itérative où une correction de Bonferroni est appliquée. Au regard des taux de détection à raison basés sur les deux critères flexibles obtenus quand $N = 800$, les algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD arrivent à détecter avec une bonne puissance les paires item-covariable affectées par du DIF de taille moyenne tel que simulé dans ces travaux (pour un questionnaire ayant une structure similaire à celle étudiée et une taille d'échantillon du même ordre). Ces algorithmes ne devraient généralement pas se tromper en détectant d'autres paires item-covariable sans DIF. Néanmoins, il faut être prudent quant à la forme de DIF mise en évidence par ces algorithmes. En effet, le DIF non homogène est rarement identifié comme tel. Cela signifie que le test effectué à l'étape 5 pour déterminer la forme du DIF manque de puissance (l'hypothèse nulle de DIF homogène n'étant pas assez souvent rejetée lorsque le DIF simulé était non homogène). L'identification correcte de la forme de DIF nécessite probablement une taille d'échantillon plus importante que celles considérées.

L'approche par pénalisation (PCM-Lasso) a quant à elle présenté des taux de détection à tort très élevés parmi les scénarios sans DIF. En effet, cette procédure détectait par erreur du DIF pour au moins une paire item-covariable dans près de la moitié des jeux de données simulés sans DIF, quel que soit le scénario considéré. Après des investigations supplémentaires parmi les jeux de données où du DIF a été détecté à tort, il s'est avéré que la procédure PCM-Lasso identifiait du DIF sur en moyenne une paire item-covariable. Cet aspect de PCM-Lasso est également mis en évidence par les écarts importants entre le taux de détection avec le critère flexible et le taux de détection avec le critère le plus flexible observés parmi les scénarios avec du DIF. Plus précisément, PCM-Lasso avait tendance à détecter à tort d'autres paires item-covariable en plus de celles réellement affectées par du DIF. Néanmoins, la taille estimée de ces effets détectés à tort restait généralement faible. Il faut également être prudent quant à la forme du DIF mise en évidence par cette approche. En effet, dans l'étude de simulation, PCM-Lasso concluait quasi systématiquement à la présence de DIF non homogène (quelle que soit la forme du DIF simulé). En s'intéressant aux estimations des paramètres de DIF parmi les scénarios avec DIF homogène, on a pu remarquer que dans certains jeux de données, l'estimation des paramètres de DIF étaient très proches (par exemple, $\gamma_{21}^{(C_1)} = 0.16, \gamma_{22}^{(C_1)} = 0.16$ et $\gamma_{23}^{(C_1)} = 0.18$). Cela signifie que le paramètre de pénalisation a probablement été choisi juste après la séparation des *parameter paths* pour les paramètres de DIF (voir la frontière entre la zone "DIF homogène" et la zone "DIF non homogène" illustrée dans la figure 5.4, graphique a). On pourrait se demander si un changement de critère (utiliser l'AIC plutôt que le BIC par exemple) pourrait améliorer les résultats. Cependant, le BIC a été justement privilégié par Schauberger et Mair puisqu'il est plus conservateur que l'AIC [19]. Cela signifie qu'il est plus susceptible de sélectionner le modèle le plus simple (le terme de pénalité sélectionné étant généralement plus grand).

Globalement, les trois méthodes évaluées n'ont pas réussi à détecter de façon satisfaisante les items affectés par du DIF faible. De plus, leurs performances de détection à raison ont fortement diminué lorsque la taille d'échantillon N passait de 800 à 400. Ces résultats étaient attendus. Le premier point ne représente peut-être pas un problème majeur, un DIF faible n'entrainant

potentiellement pas de biais de mesure substantiel au niveau de l'échelle [180]. En revanche, les performances modérées observées quand $N = 400$ sont problématiques, les effectifs des études en santé étant généralement modestes.

Parmi les jeux de données satisfaisant le critère le plus flexible, les estimations des paramètres de DIF étaient généralement proches des vraies valeurs simulées pour les algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD. Ces estimations étaient en revanche toujours sous-estimées par PCM-Lasso en raison de la pénalisation qui les "rétrécissait vers 0". Cette sous-estimation a déjà été mise en évidence dans des études de simulation antérieures menées sur cette méthode [19, 181]. Tutz et Schauberger ont indiqué que le biais introduit par la pénalisation des paramètres pourrait être éliminé en estimant un "modèle final" non pénalisé, qui inclurait uniquement les paramètres de DIF des paires identifiées comme étant affectées par du DIF. Cette "ré-estimation" entraînait néanmoins mécaniquement une augmentation des estimations des paramètres de DIF associées aux paires identifiées à tort par la procédure [181].

Enfin, les trois méthodes ont permis d'obtenir une estimation globalement non biaisée de l'effet des covariables sur le niveau moyen de la variable latente. Les estimations les plus biaisées ont été obtenues pour la covariable C_1, parmi les scénarios où cette covariable était la seule à induire du DIF sur la moitié des items (configuration n°3, $J = 4$ items). Comme l'ont démontré Rouquette *et al.* dans une étude de simulation, une telle configuration pourrait induire un biais important si le DIF est ignoré [180], or les trois procédures ont justement présenté des performances de détection du DIF plus faibles parmi ces scénarios, expliquant ainsi le biais observé. Ces scénarios sont néanmoins discutables : on peut se demander si l'on peut encore parler de DIF lorsque plus de la moitié des items sont affectés par du DIF induit par une même covariable. En effet, le DIF a été conceptualisé comme une différence dans la probabilité de répondre favorablement à certains items après avoir "contrôlé" le niveau de la variable latente (c'est-à-dire à niveau de variable latente égal par ailleurs) [15]. Si une grande proportion d'items est affectée par du "DIF", le questionnaire pourrait mesurer des construits différents parmi les groupes comparés, et cela ne ferait donc pas sens de "contrôler" le niveau de la variable latente (le construit ciblé par

le questionnaire n'étant pas conceptualisé de la même façon par les différents groupes) [175].

Limites et perspectives

Plusieurs limites peuvent être mentionnées concernant ces travaux. Tout d'abord, nous avons considéré des situations simples où les deux covariables avaient des effectifs bien équilibrés et n'impactaient pas le niveau de la variable latente. Afin de mieux appréhender les performances des trois méthodes évaluées, il serait intéressant d'étendre le cadre de simulation en considérant des covariables ayant des effectifs moins équilibrés et des covariables ayant un impact sur le niveau de la variable. Au vu de nos premiers résultats sur la taille de l'échantillon, on peut notamment s'attendre à ce que les performances des méthodes soient moindres dans le cas de covariables à effectifs déséquilibrés. En ce qui concerne l'effet des covariables sur la variable latente, plusieurs auteurs ayant évalué des méthodes de détection du DIF (item-par-item sur la base de tests statistiques ou par pénalisation) ont indiqué que la présence d'un effet des covariables sur le niveau de la variable latente n'influençait pas les résultats [19, 20, 179].

En outre, tous les jeux de données étaient complets : il n'y avait aucune donnée manquante, que ce soit pour les réponses aux items ou les covariables. Une étude de simulation avec des jeux de données incomplets pourrait être utile pour évaluer les performances des trois méthodes dans une situation plus représentative des données réelles. Dans le cas de données manquantes (complètement) au hasard, les estimations devraient être asymptotiquement sans biais (les paramètres étant estimés par maximum de vraisemblance marginale). Une perte de performance est néanmoins attendue en raison d'une perte de précision dans les estimations.

Nous avons choisi d'évaluer les trois procédures en calculant pour chaque scénario la proportion de jeux de données satisfaisant différents critères. Pour les scénarios sans DIF, le critère correspondait au fait de ne pas identifier de DIF. Pour les scénarios avec DIF, plusieurs critères ont été considérés. Le moins exigeant était de retrouver au moins les bonnes paires item-covariable pour lesquelles du DIF avait été simulé. Le plus exigeant requérait quant à lui que la procédure ait retrouvé exactement ce qui avait été simulé. Ces critères permettent donc de déterminer dans

quelle mesure chacune des méthodes a été capable de retrouver ce qui a été simulé. Ces critères de détection à raison et à tort tels que définis dans nos travaux ne sont pas "usuels" dans la littérature sur la détection du DIF. En effet, plusieurs études de simulations se sont plutôt intéressées aux taux de faux positifs (*false positive rates*, FPR) et de vrais positifs (*true positive rates*, TPR), définis par :

- FPR = Rapport entre le nombre de paires identifiées comme étant affectées par du DIF **à tort** et le nombre de paires non affectées par du DIF

$$\text{FPR} = \frac{\#(\text{paires pour lesquelles on détecte du DIF à tort})}{\#(\text{paires simulées pour ne pas être affectées par du DIF})}$$

- TPR = Rapport entre le nombre de paires identifiées comme étant affectées par du DIF **à raison** et le nombre de paires effectivement affectées par du DIF

$$\text{TPR} = \frac{\#(\text{paires pour lesquelles on détecte du DIF à raison})}{\#(\text{paires simulées pour être affectées par du DIF})}$$

Ces taux sont alors calculés pour chaque jeu de données puis moyennés au niveau du scénario. Les taux moyens de faux positifs sont ensuite comparés à 5%, mais cette comparaison est discutable. Par exemple, si l'on s'intéresse aux scénarios sans DIF de notre étude de simulation et que l'on calcule les taux de faux positifs moyens (tableau 5.11), on remarque que la procédure PCM-Lasso présente des taux de faux positifs moyennés sur l'ensemble des réplications légèrement supérieurs à 5% alors que la procédure concluait à la présence de DIF pour la moitié des réplications de ce scénario. On aurait donc conclu à des résultats globalement acceptables avec les taux de faux positifs moyens, alors que le taux de détection à tort utilisé dans notre étude mène plutôt à la conclusion d'une méthode peu adéquate.

TABLEAU 5.11 – Taux de détection à tort du DIF (%DIF détecté) et taux de faux positifs (FPR) moyens pour les scénarios sans DIF simulé.

			ROSALI-DIF FORWARD		**ROSALI-DIF BACKWARD**		**PCM-Lasso**	
N	J	Corr	%DIF détecté	FPR moyen	%DIF détecté	FPR moyen	%DIF détecté	FPR moyen
400	4	Non	4%	1%	4%	1%	50%	7%
400	4	Oui	3%	1%	3%	1%	50%	7%
400	7	Non	6%	1%	6%	1%	44%	4%
400	7	Oui	3%	<1%	4%	1%	54%	4%
800	4	Non	4%	1%	4%	1%	46%	6%
800	4	Oui	3%	1%	4%	1%	47%	6%
800	7	Non	5%	1%	5%	1%	48%	4%
800	7	Oui	3%	<1%	3%	<1%	46%	4%

Notes :
%DIF Détecté : *Proportion de jeux de données où du DIF a été détecté à tort sur au moins une paire item-covariable à la fin de la procédure*
FPR moyen : *Taux de faux positifs moyens*
N : *Taille de l'échantillon*
J : *Nombre d'items*
Corr : *Corrélation entre C_1 et C_2*

En ce qui concerne les méthodes de détection elles-mêmes, plusieurs limites peuvent également être citées. Premièrement, ni les extensions de ROSALI ni l'approche PCM-Lasso ne permettent de considérer que l'effet DIF d'une covariable peut dépendre du niveau d'une autre covariable. Pour tenir compte d'un tel phénomène, il serait nécessaire d'introduire un terme d'interaction entre les covariables. Cependant, des développements supplémentaires sont nécessaires pour savoir comment traiter une telle interaction. On peut noter que la méthode DIF-IFT [71] modélise systématiquement une telle interaction du fait de sa philosophie de partitionnement récursif, mais cette méthode n'a pas été considérée dans ces travaux, car son implémentation ne permet pas (à notre connaissance) d'estimer l'effet des covariables sur le niveau de la variable latente. Chun et al. [22] s'y étaient aussi intéressés avec leur méthode MIMIC, mais leur méthode avait des difficultés pour identifier la source du DIF.

Toutes les méthodes mentionnées sont basées uniquement sur des résultats statistiques. Il faut donc garder en mémoire que dès lors que l'effectif est assez grand, du DIF pourra être mis en évidence. Il est donc essentiel de s'assurer que les effets décelés sont pertinents et importants. On peut notamment se demander si les effets mis en évidence font sens ? S'ils sont interprétables ? Justifiables d'un point de vue clinique ou culturel ? La recherche du DIF ne doit pas s'arrêter à une conclusion binaire indiquant uniquement s'il y a, ou non, du DIF. En outre, il pourrait être intéressant d'ajouter des connaissances cliniques *a priori* sur les items pour lesquels un fonctionnement différentiel est probable (ou en tout cas ferait sens). Cela permettrait de ne pas tester l'ensemble des items et de ne pas se fier uniquement à des critères statistiques.

Pour conclure, ces travaux ont permis de fournir des données empiriques sur les performances de détection du DIF de trois méthodes : ROSALI-DIF FORWARD, ROSALI-DIF BACKWARD et PCM-Lasso. Ils ont aussi soulevé des questions méthodologiques et conceptuelles concernant la détection du DIF. Dans notre cadre de simulation, les méthodes itératives (ROSALI DIF FORWARD et ROSALI-DIF BACKWARD) ont semblé plus performantes que la méthode basée sur la pénalisation de vraisemblance. Des développements supplémentaires sont néanmoins nécessaires afin, à terme, de mieux comprendre les déterminants du manque d'invariance de la mesure lors de l'analyse des données rapportées par les patients.

5.8 Analyses *post hoc*

Performances de PCM-Lasso pour différents paramètres de pénalisation

L'idée de cette analyse *post hoc* était d'étudier les performances de la procédure PCM-Lasso pour différents paramètres de pénalisation λ fixés à l'avance (et non pas sélectionnés pour chaque réplication en cherchant le modèle minimisant le BIC). L'objectif était de déterminer si on pouvait se rapprocher des performances des algorithmes ROSALI-DIF FORWARD et BACKWARD avec un paramètre de pénalisation différent. Pour ce faire, nous avons étudié un scénario issu de l'étude de simulation présenté précédemment. Il s'agit du scénario avec 800 individus, 4 items et où les covariables C_1 et C_2 induisent respectivement du DIF homogène de taille moyenne sur les items 2 et 3. L'analyse réalisée sur ce scénario est représentée graphiquement dans la figure 5.12.

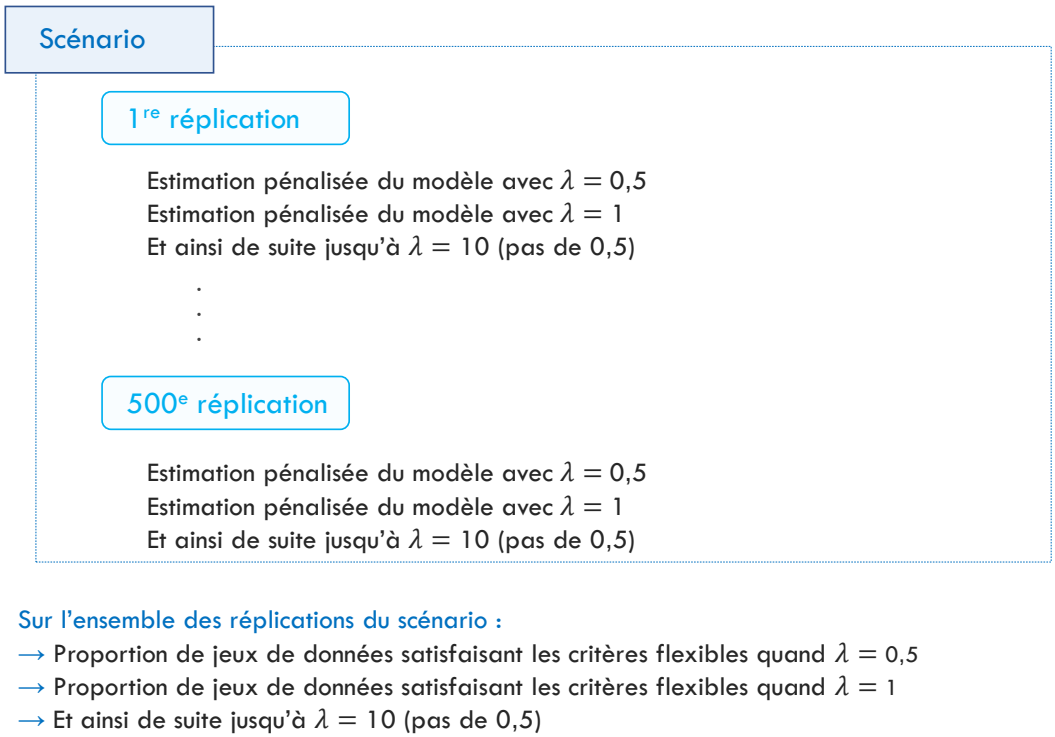


FIGURE 5.12 – Analyse *post hoc* n°1

Pour résumer, les taux de détection à raison du DIF (critères flexibles) ont été calculés sur ce scénario pour différents paramètres de pénalisation (allant de λ = 0,5 à λ = 10 avec un pas de 0,5).

Les résultats obtenus sont donnés dans la figure 5.13, où les taux de détection à raison à raison du DIF associées à la procédure PCM-Lasso sont représentées en fonction des valeurs considérées pour le paramètre de pénalisation λ. Ces résultats semblent indiquer que même avec une autre méthode de sélection du paramètre de pénalisation, les performances de PCM-Lasso resteraient inférieures à celles des algorithmes ROSAI-DIF.

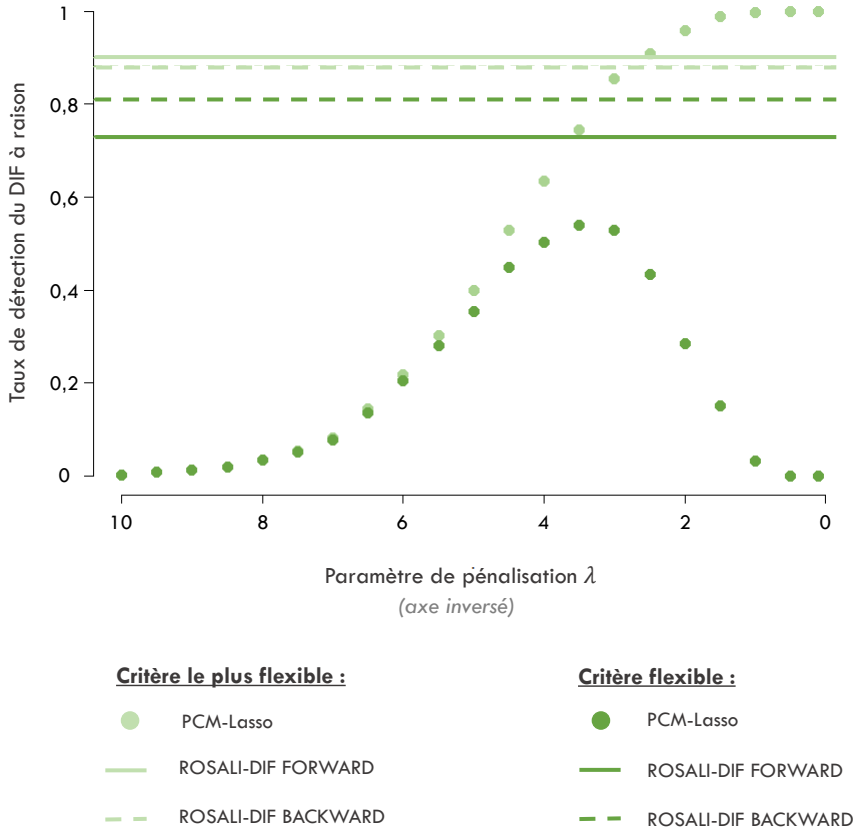


FIGURE 5.13 – Taux de détection du DIF à raison pour la procédure PCM-Lasso avec différents paramètres de pénalisation

Notes :
Critère le plus flexible *: indique si la procédure a détecté du DIF sur au moins les bonnes paires item-covariable (parmi d'autres)*
Critère flexible *: indique si la procédure a détecté du DIF uniquement sur les bonnes paires item-covariable*

Estimation de l'effet des covariables sur le niveau moyen de la variable latente

Afin de déterminer pourquoi les estimations des paramètres β_1 et β_2 étaient moins dispersées avec la méthode PCM-Lasso qu'avec les deux algorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD, des analyses complémentaires ont été réalisées. Théoriquement, avec PCM-Lasso, l'estimation des paramètres β_1 et β_2 ne devraient pas être impactés par la pénalisation, ces termes n'intervenant pas dans le terme de pénalisation de type lasso décrit dans l'équation 5.10. Néanmoins, la différence d'erreurs standards empiriques observée entre PCM-Lasso et les deux autres méthodes pourrait potentiellement s'expliquer par une régularisation de ces termes alors qu'ils n'auraient pas dû être pénalisés.

Comme les scénarios explorés ne permettent pas d'investiguer cette hypothèse (les effets des covariables sur la variable latente ayant été fixé à 0), un scénario supplémentaire a été réalisé *a posteriori*. Ses caractéristiques sont décrites dans le tableau 5.12. Comme précédemment, ce scénario a été répliqué 500 fois.

TABLEAU 5.12 – Description du scénario supplémentaire étudié *a posteriori*

Structure du questionnaire et échantillon	
Nombre d'items	$J = 4$
Nombre de modalités de réponse	$M = 4$ modalités de réponse
Taille de l'échantillon N	800 individus simulés
Variable latente Θ	
Moyenne μ , Variance σ^2	$\mu = 0$, $\sigma^2 = 1$
Effet des covariables sur la variable latente	$\beta_1 = -0{,}2$ $\beta_2 = -0{,}2$
DIF	
Covariables à l'origine du DIF	C_1 et C_2, non corrélées
Items affectés par du DIF	Item 2 (DIF induit par C_1) Item 3 (DIF induit par C_2)
Forme du DIF	Homogène
Taille du DIF	Moyenne

Les trois méthodes de détection du DIF ont été appliquées sur l'ensemble des jeux de données simulés. Pour ROSALI-DIF FORWARD et PCM-Lasso, la procédure de détection du DIF s'est terminée normalement pour les 500 jeux de données. L'algorithme ROSALI-DIF BACKWARD s'est en revanche arrêté avant la fin de la procédure pour quatre jeux de données (l'algorithme n'ayant pas réussi à identifier au moins un item *anchor* pour une des deux covariables, le modèle final n'était donc pas identifiable). Pour les 99% de jeux de données restants, l'algorithme s'est terminé normalement. Les taux de détection à raison du DIF sont décrits dans le tableau 5.13 et les estimations obtenues pour β_1 et β_2 (ajustées sur le DIF mis en évidence, le cas échéant) sont représentées dans la figure 5.14 par des *violin plots* pour chacune des trois méthodes étudiées.

TABLEAU 5.13 – Taux de détection à raison du DIF pour le scénario supplémentaire étudié *a posteriori*

	%Plus flexible	**%Flexible**	**%Parfait**
ROSALI-DIF FORWARD (n_{sim} = 500)	92%	82%	75%
ROSALI-DIF BACKWARD (n_{sim} = 496)	88%	84%	77%
PCM-Lasso (n_{sim} = 500)	75%	40%	1%

Notes :
$\boldsymbol{n_{sim}}$: *Nombre de réplications pour lesquelles la procédure s'est terminée normalement*
%Plus flexible : *Proportion de jeux de données où la procédure a détecté du DIF sur au moins les bonnes paires item-covariable (parmi d'autres)*
%Flexible : *Proportion de jeux de données où la procédure a détecté du DIF uniquement sur les bonnes paires item-covariable*
%Parfait : *Proportion de jeux de données où la procédure a retrouvé exactement le DIF simulé (bonnes paires uniquement et bonne forme de DIF)*

Comme pour l'étude de simulation, les estimations obtenues avec PCM-Lasso sont moins dispersées que celles obtenues avec les deux autres algorithmes. Si l'on compare le biais obtenu sur ce nouveau scénario à celui du scénario ayant les mêmes caractéristiques, mais où $\beta_1 = \beta_2 = 0$, on peut remarquer que le biais obtenu avec PCM-Lasso a doublé en valeur absolue avec ce nouveau scénario (0.07 contre -0.03, voir la figure 5.15). Les estimations de β_1 et β_2 sont en fait centrées sur -0.13 au lieu d'être centrées sur -0.2 comme c'est le cas pour ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD. Cette sous-estimation des paramètres et la plus

faible erreur standard empirique sont des éléments concordants en faveur de l'hypothèse d'une estimation pénalisée des paramètres β_1 et β_2.

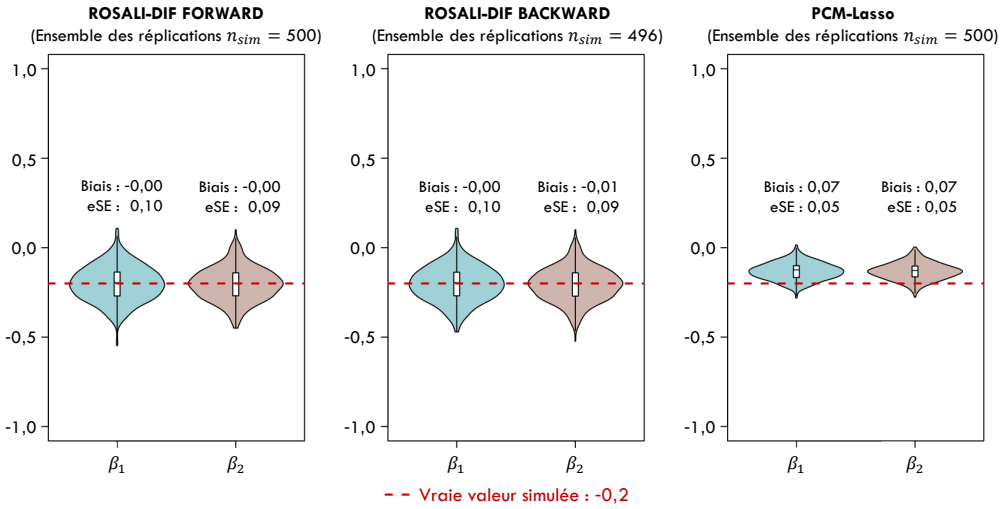


FIGURE 5.14 – *Violin plot* des estimations de β_1 et β_2 (nouveau scénario)

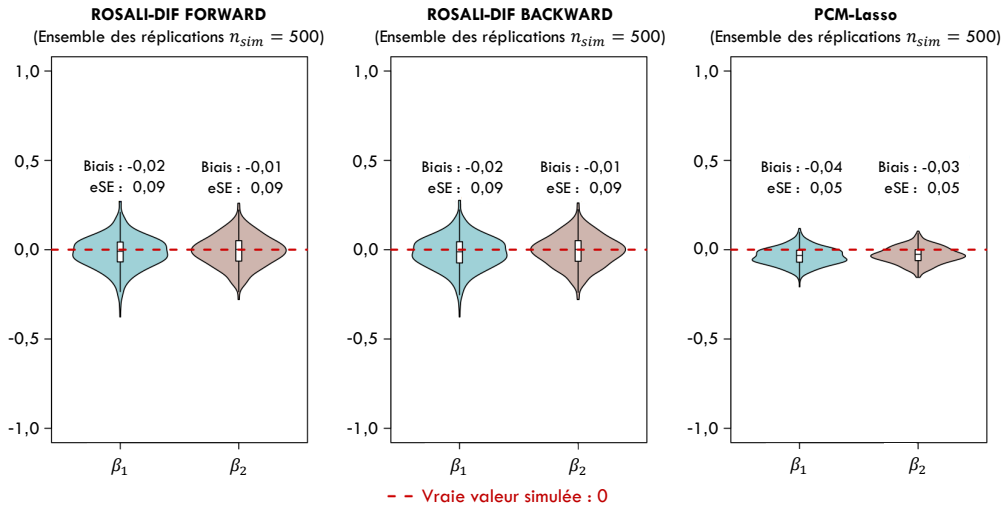


FIGURE 5.15 – *Violin plot* des estimations de β_1 et β_2 (scénario de l'étude de simulation)

Notes :

β_1 et β_2 *: Effets respectifs des covariables C_1 et C_2 sur le niveau moyen de la variable latente*

n_{sim} *: Nombre de réplications pour lesquelles la procédure s'est terminée normalement*

eSE *: Erreur standard empirique*

Chapitre 6

PTGI : Propriétés psychométriques d'une des versions françaises chez des patients atteints d'un cancer

Sommaire

6.1 Motivations et objectifs

Les travaux présentés dans ce chapitre sont centrés autour de l'étude des propriétés psychométriques d'une des versions françaises de l'inventaire du développement post-traumatique [24, 182].

Le concept de développement post-traumatique s'inscrit dans le courant de la psychologie positive et désigne les changements psychologiques positifs perçus par certains individus suite à la confrontation avec un événement de vie défiant hautement leurs ressources [24]. Ce concept nous a semblé intéressant à étudier, car potentiellement lié au *response shift*. En effet, ces deux phénomènes font suite à un "catalyseur" (un événement de vie traumatique pour le développement post-traumatique et un événement de vie ou de santé saillant pour le *response shift*) et partagent potentiellement des mécanismes communs.

La première partie de ce chapitre présente plus en détail le concept de développement post-traumatique et l'outil principal qui permet de l'évaluer. Ce chapitre se poursuit ensuite avec l'étude des propriétés psychométriques d'une des versions françaises de l'inventaire de développement post-traumatique [182]. Pour terminer, un bilan est dressé en toute fin de chapitre.

6.2 Le développement post-traumatique : définition et modèle

Le développement post-traumatique (ou croissance post-traumatique) désigne l'ensemble des "changements psychologiques positifs résultant de la confrontation, de la lutte avec un événement de vie défiant hautement les ressources de l'individu" [24, 183]. Le terme de développement post-traumatique et sa définition sont apparus au cours des années 1990, dans les travaux de Tedeschi et Calhoun [24, 184]. Cette définition conceptualise l'idée que les individus confrontés à une situation difficile, à laquelle ils n'étaient pas préparés, peuvent percevoir des changements psychologiques positifs comme [24, 184] :

– Des changements dans la perception de soi, avec le développement d'un sentiment de force personnelle et de confiance en soi ;

– Des changements dans la philosophie de la vie, avec la redéfinition des priorités et de la direction pour la vie, le développement d'une certaine spiritualité et une plus grande appréciation de la vie ;

– Des changements dans les relations avec les autres, qui peuvent devenir plus profondes et plus investies.

Il est important de souligner que la théorie du développement post-traumatique ne nie pas l'existence de changements négatifs qui peuvent survenir à la suite d'un événement traumatique. Au contraire, Tedeschi et Calhoun énoncent clairement le "paradoxe" au cœur de leur théorie, en indiquant que du bien peut découler de la souffrance [185]. On retrouve notamment cette ambivalence dans le modèle du développement post-traumatique représenté dans la figure 6.1, dont la dernière version date de 2018 [186]. C'est ce modèle qui va être résumé ici.

Avant d'être confronté à un événement de vie grave, chaque individu a ses propres croyances fondamentales et schémas cognitifs [187] préexistants, c'est-à-dire sa façon personnelle de percevoir et d'évaluer les autres, le monde en général, mais également de se percevoir et de s'évaluer soi-même. Ces croyances fondamentales et ces schémas sont des "lunettes" au travers desquelles on perçoit la vie.

Face à un événement de vie très grave et profondément désorganisateur, il est possible que les objectifs, les croyances fondamentales et les schémas cognitifs de l'individu soient remises en question, voir s'effondrent, engendrant une détresse émotionnelle profonde. Peuvent s'ensuivre des ruminations sur le plan cognitif et une tentative de régulation sur le plan émotionnel. L'individu devra, dans un premier temps, faire face à des pensées intrusives et automatiques à propos de l'événement vécu et de sa place dans l'existence. Ces pensées deviendront petit à petit délibérées, réfléchies et constructives, à mesure que l'individu revoit ses buts, atténue sa détresse émotionnelle et s'ajuste à cette situation de stress en mettant en place des stratégies de *coping*. En effet, l'individu va être progressivement amené à abandonner ses anciens objectifs de vie (qui ne sont désormais plus atteignables face à la nouvelle réalité de sa vie post-trauma) et les remplacer par la construction de nouveaux. Cette redéfinition des objectifs de vie va participer

à diminuer l'intensité de la détresse émotionnelle, permettant ainsi d'initier une démarche de réflexion plus constructive.

Le processus qui vient d'être décrit est long et graduel, car il implique un remaniement en profondeur des croyances fondamentales et des schémas cognitifs des individus, amenant les individus à porter un nouveau regard sur le monde, sur les autres et sur eux-mêmes. Selon Tedeschi *et al.*, ce serait grâce à la refonte des buts et des croyances fondamentales que l'individu pourra accepter sa situation et percevoir des changements positifs suite à l'expérience vécue [186].

L'expression des émotions et le partage social jouent un rôle capital dans ce modèle. En effet, c'est en partageant le vécu présent, les préoccupations passées et présentes et les inquiétudes pour le futur que l'individu va pouvoir clarifier sa pensée. Verbaliser sa pensée peut aider l'individu à développer des pensées alternatives en lien avec la situation vécue, et contribuer à assouplir ses croyances fondamentales. Il convient néanmoins de rappeler que le développement post-traumatique n'est pas universel ; face à un même événement, les individus n'expérimenteront pas tous du développement post-traumatique.

Dans le modèle, le développement post-traumatique est présenté comme un *outcome* (une issue), mais il y a en fait un débat récurrent dans la littérature sur la nature du développement post-traumatique : est-ce un *outcome* (une issue avec de réels changements intégrés dans la définition de soi) ou un processus (forme de *coping* cognitif visant à aider l'individu à faire face à l'événement qui défie ses croyances et ses valeurs) ? D'après Tedeschi et Calhoun, le développement post-traumatique est à la fois l'un et l'autre, successivement, à mesure que le temps s'écoule depuis la survenue de l'événement [186].

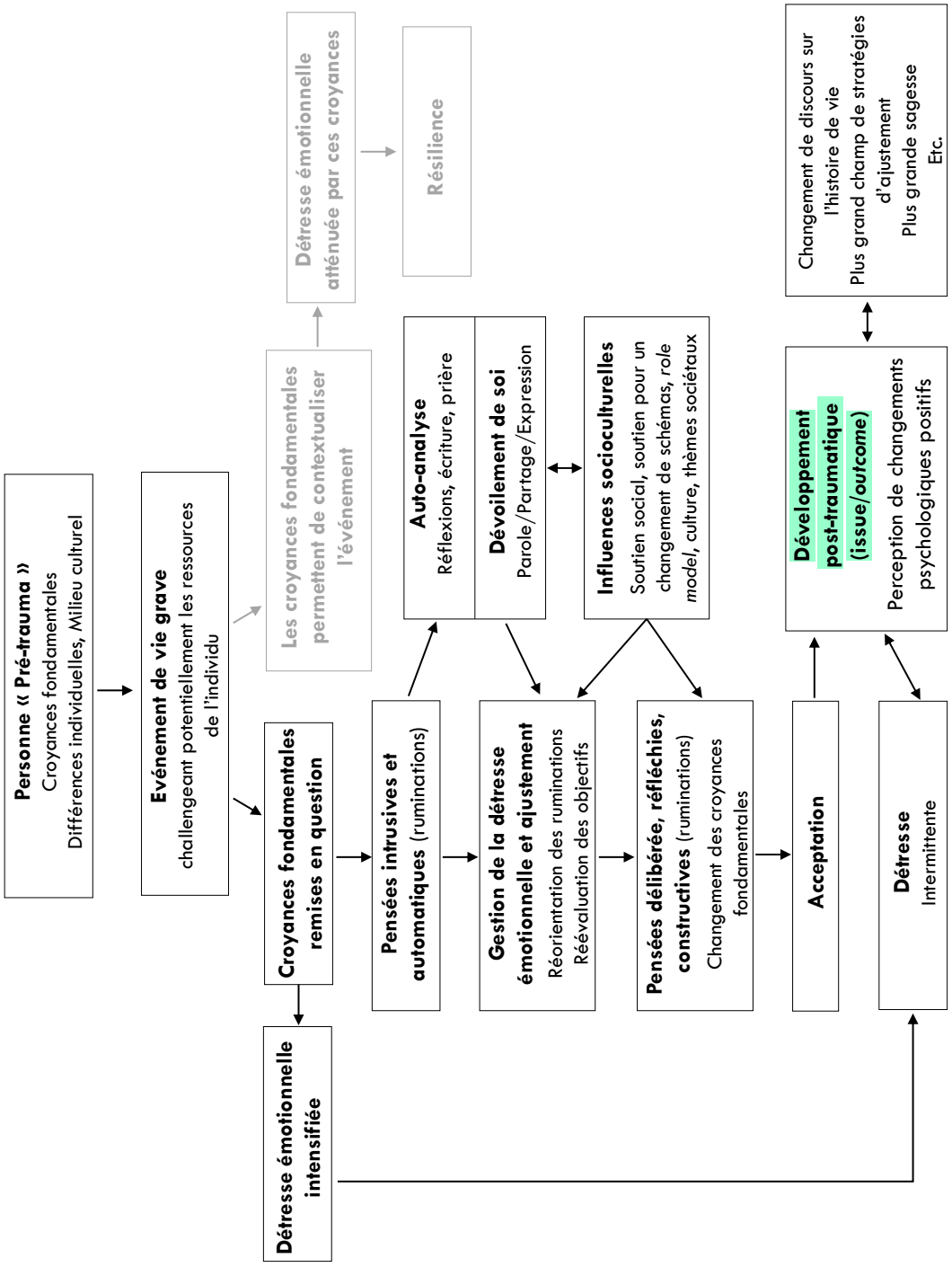


FIGURE 6.1 – Modèle révisé du développement post-traumatique
Source : Adapté de Tedeschi, Shakespeare-Finch, Taku, Calhoun. Posttraumatic Growth : Theory, Research, and Applications (2018). Routledge

Initialement, les travaux s'intéressant au développement post-traumatique se sont concentrés sur des événements comme le deuil, les violences interpersonnelles, les catastrophe naturelles, les traumas de guerre ou encore les accidents de transport. Ce concept s'est ensuite étendu à la psychologie de la santé, avec la recherche de développement post-traumatique chez des patients atteints de maladies graves. Ces maladies sont considérées comme source de traumatisme puisqu'elles peuvent remettre en question la survie de l'individu ou peuvent être caractérisées par une chronicité [188]. On s'intéresse aujourd'hui au développement post-traumatique dans de nombreuses pathologies comme le cancer, l'infection par le virus de l'immunodéficience humaine, l'insuffisance rénale chronique, etc. [186].

Le développement post-traumatique et le *response shift* sont des phénomènes d'intérêt dans certains champs de santé, comme les thérapies comportementales et cognitives ou les soins de réadaptation [58, 186, 189]. En effet, on y cherche parfois à accompagner ou promouvoir une croissance post-traumatique ou un "*response shift* positif" (c.-à-d., un changement positif dans la façon de penser à ses normes de mesure interne, ses priorités ou sa conception de la santé). Par exemple, dans le champ des soins de réadaptation, Mayo a rappelé que les thérapeutes cherchent à modifier les schémas de pensées des patients, afin qu'ils portent un autre regard sur leur situation, se concentrent sur ce qu'ils peuvent faire, plutôt que ce qu'ils ne peuvent pas faire [62]. L'objectif de cet accompagnement est donc d'aider les patients à réorganiser leurs valeurs et leurs objectifs, de façon à ce que les objectifs réalisables gagnent en valeur et les objectifs irréalisables deviennent moins importants (notion que l'on retrouve à la fois dans la théorie du *response shift* en lien avec l'adaptation et dans la théorie du développement post-traumatique).

Le développement post-traumatique et le *response-shift* pourraient en fait partager des mécanismes comportementaux, affectifs et cognitifs proches ou communs. Vanier *et al.* [18] ont notamment suggéré que le développement post-traumatique pourrait être une cause possible du *response shift*. L'étude des liens entre ces phénomènes pourraient être intéressante pour documenter cette hypothèse. On pourrait notamment se questionner sur la temporalité liant ces deux phénomènes et s'intéresser aux formes de *response shift* qui pourraient (ou non) être retrou-

vées chez les individus qui expérimentent du développement post-traumatique. À terme, l'étude conjointe de ces phénomènes pourrait permettre d'identifier des leviers d'action pour accompagner au mieux les patients. Pour pouvoir mener à bien ce type de recherche, il est néanmoins nécessaire d'avoir un outil de mesure en français du développement post-traumatique qui soit fiable et valide. L'objectif des travaux présentés dans ce chapitre visaient donc à répondre à ce besoin.

6.3 Mesurer le développement post-traumatique

L'outil le plus répandu pour évaluer le développement post-traumatique est l'inventaire du développement post-traumatique (*Posttraumatic Growth Inventory*, PTGI) proposé par Tedeschi et Calhoun en 1996 en langue anglaise [24]. Ce questionnaire est composé de 21 items qui ont été répartis en cinq dimensions à la suite d'une analyse exploratoire : "Relation aux autres" (sept items), "Nouvelles opportunités" (cinq items), "Force Personnelle" (quatre items), "Changement spirituel" (deux items) et "Appréciation de la vie" (trois items). Ces cinq dimensions sont directement en lien avec les trois formes de développement initialement décrites [184] (c.-à-d., changements dans la perception de soi, dans la philosophie de la vie et dans les relations avec les autres). Elles permettent en outre d'obtenir une image plus fine des types de changements qui peuvent être vécus.

Ce questionnaire a été utilisé pour évaluer le développement post-traumatique à la suite d'événements traumatiques multiples et variés. Il a également été traduit dans de nombreuses langues (au moins 25) avec parfois plusieurs versions pour une même langue. C'est notamment le cas pour le français. Durant ces vingt dernière années, une certaine instabilité structurelle est apparue, avec des études proposant une structure à une, deux, trois, quatre ou cinq dimensions. Cette instabilité structurelle pourrait en partie s'expliquer par des différences culturelles. Garrido-Hernansaiz *et al.* [190] ont par exemple suggéré que les cinq dimensions proposées par Tedeschi et Calhoun [24] avaient tendance à apparaître parmi les cultures individualistes, alors que dans les cultures plus collectivistes, les dimensions qui reflètent le développement post-

traumatique au niveau individuel (c.-à-d., "Nouvelles opportunités", "Appréciation de la vie" et "Force personnelle") formaient une seule dimension.

En France, au moins trois versions sont actuellement utilisées, elles sont décrites ci-dessous :

- **Version 1 :** Il s'agit de la traduction originelle réalisée par Lelorain *et al.* [182]. Dans la littérature, aucune information n'est disponible sur le processus de traduction mis en œuvre. Après avoir contacté ces auteures, nous avons appris qu'elles avaient traduit elles-mêmes le questionnaire en français, et qu'un enseignant d'anglais (natif anglais) avait effectué une rétro-traduction. Cependant, nous ne savons pas si des divergences avaient été mises en évidence, et comment elles avaient été prises en compte le cas échéant. Dans cette version, il était demandé aux patients d'indiquer à quel point leur cancer avait engendré les changements mentionnés par les items. Chaque item comportait six modalités de réponse allant de 0 = "Pas du tout" à 5 = "Totalement". Cette version a ensuite été généralisée à d'autres événements que le cancer.
- **Version 2 :** Suite aux travaux de Lelorain *et al.* [182] une version révisée est apparue dans les protocoles d'études françaises à partir de 2010. Cette version révisée présente de légères adaptations de formulation par rapport à la traduction originelle : au total, des modifications de formulation ont été observées pour quatre items (sur les 21 items au total).
- **Version 3 :** Une autre traduction française de l'inventaire du développement post-traumatique a été effectuée par Cadell *et al.* [191]. Pour traduire le questionnaire de l'anglais au français, ces auteurs ont fait appel à un traducteur professionnel. Une rétro-traduction indépendante a ensuite été réalisée par un autre traducteur et aucune différence n'a été mise en évidence. Le contenu des items a été validé de manière indépendante par un locuteur français dont c'était la langue maternelle. Cadell *et al.* ont fait le choix de retraduire l'inventaire du développement post-traumatique en français parce qu'ils n'ont pas trouvé de version publiée de la traduction de Lelorain *et al.* [182]. La formulation française des items qu'ils proposent est disponible dans leur manuscrit, la traduction des modalités de réponse n'est en revanche pas indiquée [191].

La formulation des items pour chacune de ces versions est donnée dans le tableau 6.1 à la fin de cette section

Les données sur les propriétés psychométriques des versions 1 et 2 ne sont pas suffisantes pour statuer sur leur validité de construit. Pour la version 1, Lelorain *et al.* ont évoqué quelques propriétés psychométriques, obtenues à partir d'un échantillon composé de 307 femmes françaises en rémission d'un cancer du sein depuis au moins cinq ans [182]. Ces auteures ont rapporté qu'une analyse factorielle confirmatoire hiérarchique (basée sur la structure initiale à cinq dimensions proposée par Tedeschi et Calhoun [24]) ne présentait pas un bon ajustement, mais que les cinq dimensions présentaient néanmoins une bonne consistance interne (mesurée par le coefficient alpha de Cronbach). Il en allait de même pour l'ensemble des 21 items. Les scores aux cinq dimensions et le score total étaient ensuite utilisés dans le reste de leur article. Pour la version 2, aucune information n'était disponible avant ces travaux.

Plus d'informations sont disponibles sur la version 3 (celle de Cadell *et al.* [191]). En effet, ces auteurs ont validé leur version française de l'inventaire du développement post-traumatique sur un échantillon d'aidants canadiens francophones. Au sein de cet échantillon, on dénombre : 10 aidants endeuillés par le VIH/SIDA et 37 parents s'occupant d'un enfant atteint d'une maladie limitant son espérance de vie. Toutes les dimensions, à l'exception de la dimension "Force personnelle", présentaient des coefficients alpha de Cronbach supérieurs à 0,60 (le seuil utilisé par les auteurs comme preuve de fiabilité pour une étude exploratoire). Des analyses factorielles confirmatoires (CFA), réalisées séparément pour chacune des cinq dimensions du questionnaire, ont indiqué une validité convergente au niveau des items satisfaisante (les valeurs des *factor loadings* étant supérieures à 0,50). En revanche, aucune CFA multidimensionnelle combinant les cinq dimensions et leurs items n'a été réalisée (à la place les auteurs ont réalisé une CFA sur les scores aux cinq dimensions). La dimensionnalité du questionnaire (rapportée sous le terme de "validité discriminante") a également été explorée à l'aide de CFA emboitées. L'objectif était de déterminer si les cinq dimensions représentaient des construits distincts. C'était globalement le cas. Si les propriétés psychométriques de cette version sont connues sur un échantillon de

Canadiens francophones, les résultats de Cadell *et al.* [191] ne sont peut-être pas directement transposables en France. En effet, des différences culturelles et linguistiques existent entre les Français et les Canadiens francophones. Par ailleurs, la formulation des items de cette version est différente de celle des versions 1 et 2 (versions qui sont, à notre connaissance, les plus utilisées en France).

Nous avons donc étudié les propriétés psychométriques de l'une des versions françaises de l'inventaire du développement post-traumatique (la version 2, celle dérivant de la traduction de Lelorain *et al.* [182]) chez des individus atteints d'un cancer du sein ou d'un mélanome de stade précoce, deux ans après le diagnostic de leur cancer. L'objectif était de fournir des informations sur la validité de construit et la consistance interne de ce questionnaire. Cette étude est présentée dans la section suivante.

Tableau 6.1 – Versions françaises de l'inventaire du développement post-traumatique

N° Item	Version	Item
Q1	T&C	I changed my priorities about what is important in life
	V1	J'ai changé de priorités dans la vie
	V2	J'ai changé de priorités dans la vie
	V3	Mes priorités ont changé
Q2	T&C	I have a greater appreciation for the value of my own life
	V1	J'apprécie plus ma vie à sa vraie valeur
	V2	J'apprécie plus ma vie à sa vraie valeur
	V3	J'apprécie mieux la valeur de ma vie
Q3	T&C	I developed new interests
	V1	Je me suis intéressé(e) à de nouvelles choses
	V2	Je me suis intéressé(e) à de nouvelles choses
	V3	J'ai de nouveaux centres d'intérêt
Q4	T&C	I have a greater feeling of self-reliance
	V1	J'ai acquis plus de confiance en moi
	V2	J'ai acquis plus confiance en moi
	V3	Je compte davantage sur moi
Q5	T&C	I have a better understanding of spiritual matters
	V1	J'ai développé une certaine spiritualité
	V2	J'ai développé une certaine spiritualité
	V3	Je comprends mieux ce qui a trait à la spiritualité

Tableau 6.1 – Suite

N°Item	Version	Item
Q6	T&C	I more clearly see that I can count on people in times of trouble
	V1	Je vois mieux que je peux compter sur les autres en cas de problème
	V2	Je vois mieux que je peux compter sur les autres en cas de problème
	V3	Je me rends mieux compte que je peux compter sur les autres en cas de problème
Q7	T&C	I established a new path for my life
	V1	J'ai donné une nouvelle direction à ma vie
	V2	J'ai donné une nouvelle direction à ma vie
	V3	J'ai donné une nouvelle orientation à ma vie
Q8	T&C	I have a greater sense of closeness with others
	V1	Je me sens plus proche des autres
	V2	Je me sens plus proche des autres
	V3	Je me sens plus proche des autres
Q9	T&C	I have a greater willingness to express my emotions
	V1	Je suis plus enclin(e) à exprimer mes émotions
	V2	Je suis plus enclin(e) à exprimer mes émotions
	V3	J'exprime plus volontiers mes émotions
Q10	T&C	I know better that I can handle difficulties
	V1	Je suis davantage capable de gérer des situations difficiles
	V2	Je suis davantage capable de gérer des situations difficiles
	V3	Je sais davantage que je peux faire face aux difficultés
Q11	T&C	I'm able to do better things with my life
	V1	Je fais de ma vie quelque chose de meilleur
	V2	Je fais de ma vie quelque chose de meilleur
	V3	Je suis capable de faire de meilleures choses dans ma vie
Q12	T&C	I am better able to accept the way things work out
	V1	J'accepte mieux la façon dont les choses se passent
	V2	J'accepte mieux la façon dont les choses se passent
	V3	J'accepte plus facilement la tournure que prennent les événements
Q13	T&C	I can better appreciate each day
	V1	J'apprécie davantage chaque jour de ma vie
	V2	J'apprécie plus amplement chaque jour de ma vie
	V3	J'apprécie davantage le présent
Q14	T&C	New opportunities are available which wouldn't have been otherwise
	V1	De nouvelles opportunités sont apparues qui ne seraient pas apparues autrement
	V2	De nouvelles opportunités sont apparues
	V3	De nouvelles opportunités sont apparues, ce qui n'aurait pas été le cas auparavant
Q15	T&C	I have greater compassion for others
	V1	J'ai plus de compassion pour les autres
	V2	J'ai plus de compassion pour les autres
	V3	J'ai davantage de compassion pour les autres
Q16	T&C	I put more effort into my relationships
	V1	J'investis plus mes relations aux autres
	V2	J'investis plus mes relations aux autres
	V3	Je fais davantage d'efforts dans mes relations

Tableau 6.1 – Suite

N°Item	Version	Item
Q17	T&C	I'm more likely to try to change things which need changing
	V1	J'essaie davantage de changer les choses qui ont besoin d'être changées
	V2	J'essaie davantage de changer les choses qui ont besoin d'être changées
	V3	Je suis plus encliné(e) à changer ce qui doit l'être
Q18	T&C	I have a stronger religious faith
	V1	J'ai une foi religieuse plus grande
	V2	J'ai une foi religieuse plus grande
	V3	Ma foi s'est renforcée
Q19	T&C	I discovered that I'm stronger than I thought I was
	V1	J'ai découvert que je suis plus fort(e) que ce que je pensais
	V2	J'ai découvert que je suis plus fort(e) que ce que je pensais
	V3	J'ai découvert que j'étais plus fort(e) que je ne le pensais
Q20	T&C	I learned a great deal about how wonderful people are
	V1	J'ai vraiment compris à quel point les gens pouvaient être formidables
	V2	Je vois plus le bon côté des gens
	V3	J'ai appris à quel point les gens peuvent être merveilleux
Q21	T&C	I better accept needing others
	V1	J'accepte mieux le fait d'avoir besoin des autres
	V2	J'accepte mieux le fait d'avoir besoin des autres
	V3	J'accepte mieux d'avoir besoin des autres

Notes :
T&C : Formulation initiale des items proposée par Tedeschi et Calhoun
V1 : Version française 1 (traduction de Lelorain et al.)
V2 : Version française 2 (version révisée qui dérive de la traduction de Lelorain et al.)
V3 : Version française 3 (traduction de Cadell et al.)

6.4 Article : Posttraumatic Growth Inventory - Challenges with its validation among French cancer patients

Cet article a été publié dans la revue *BMC Medical Research Methodology* [192].

Une légère différence de formulation entre la version 1 et la version 2 est passée inaperçue (elle n'est pas mentionnée dans l'article). Il s'agit de la disparition de la préposition "de" dans l'énoncé de l'item n°4 (voir tableau 6.1).

Dubuy *et al.*
BMC Medical Research Methodology (2022) 22:246
https://doi.org/10.1186/s12874-022-01722-6

BMC Medical Research Methodology

RESEARCH Open Access

Posttraumatic growth inventory: challenges with its validation among French cancer patients

Yseulys Dubuy[1*], Véronique Sébille[1,2], Marianne Bourdon[1,3], Jean-Benoit Hardouin[1,2,4] and Myriam Blanchin[1]

Abstract

Background: The Posttraumatic growth inventory (PTGI) aims to assess the positive psychological changes that individuals can perceive after a traumatic life event such as a cancer diagnosis. Several French translations of the PTGI have been proposed, but comprehensive data on their psychometric properties are lacking. This study aimed to provide a more complete assessment of the psychometric properties of one of the most used PTGI translations in early-stage breast cancer and melanoma patients.

Methods: A sample of 379 patients completed the PTGI two years after their cancer diagnosis. A confirmatory analysis was first performed to determine whether the initial five-factor structure of the PTGI was adequate for this French version. As issues were identified in the translation and in the questionnaire structure, we performed an exploratory analysis to determine the most suitable structure for this questionnaire. Validity and reliability of the evidenced structured were then assessed.

Results: The exploratory analysis evidenced a four-factor structure close to the initial structure: four of the five initial domains were recovered, and items from the unrecovered domain were split into the other domains. This new structure showed good internal consistency and acceptable validity.

Conclusions: This study highlights that the process of translation and cross-cultural validation of questionnaires is crucial to obtain valid and reliable psychometric instruments. We advise French psycho-oncology researchers and psychotherapists to (i) use the revised translation of Lelorain et al. (2010) proposed in this manuscript and (ii) use the four scores newly evidenced with a grouping of two response categories.

Keywords: Posttraumatic growth, Psychometric properties, Psycho-oncology, Breast cancer, Melanoma

Introduction

A cancer diagnosis is a traumatic event with many consequences on patients' life and health in both the short and long terms. The changes caused by cancer in terms of socio-economic and psychological aspects are numerous, with deteriorated health-related quality of life (HRQoL) and well-being, the occurrence of mood disorders such as anxiety and depression, and loss of income due to a job loss or a working time reduction [1–6]. Patients can nonetheless also experience positive psychological changes after cancer. For instance, Sears et al. demonstrated in a qualitative interview-based study that 83% of women diagnosed with early-stage breast cancer found at least one benefit from their experience with cancer [7]. Experiencing positive psychological changes in the

*Correspondence: Yseulys.Dubuy@univ-nantes.fr

[1] Nantes Université, Université de Tours, INSERM, MethodS in Patients-Centered Outcomes and HEalth Research, SPHERE, F-44000 Nantes, France
Full list of author information is available at the end of the article

aftermath of a struggle with highly challenging life circumstances has been described in the 1990s by Tedeschi and Calhoun as *Posttraumatic growth* (PTG) [8, 9]. PTG can arise, with various intensity, after many different types of events such as war, bereavement, natural disaster, accident, assault or cancer diagnosis [10].

Three forms of positive change have been initially described [9]. First, PTG can manifest itself through perceived changes in self, with increased self-reliance, self-confidence and sense of strength to cope with difficult situations in the future. In addition, individuals facing traumatic events can also change their philosophy of life by redefining their life priorities, growing in spirituality or better appreciating their life. Finally, the experience of trauma can also positively modify interpersonal relationships. Indeed, traumatic events can strengthen bounds between individuals and change the way people interact with each other.

The most widely used measurement tool for assessing PTG is the Posttraumatic Growth Inventory (PTGI) developed by Tedeschi and Calhoun [9]. The PTGI is a self-reported questionnaire composed of 21 items organized into five domains associated with the three abovementioned forms of positive change. Of note, when the questionnaire was developed, only these three abovementioned forms of positive change were well identified in the literature; the five domains emerged from the evaluation of the psychometric properties of the PTGI and covered these three expected forms of positive change. Besides, these five domains gave a more refined picture of perceived positive changes. This questionnaire is available in at least 25 languages [10], and its psychometric properties have been evaluated in many populations with various types of traumas. These studies have shown some structural instability, with some authors finding one, two, three, four, or five domains. In the French language, the PTGI has been translated twice. The first translation was performed by Lelorain et al. [11], and the second one was realized by Cadell et al. [12]. Yet, to date, at least three French versions derived from these translations are currently used in France, and comprehensive data regarding their psychometric properties are lacking, notably in cancer patients.

In France, PTG is a growing field in research and clinical care related to life after a trauma [13] such as a cancer diagnosis. Ensuring the validity and reliability of the PTGI questionnaire is essential for researchers to get more insight into the experience of positive change following a cancer diagnosis in clinical research studies. It is also important in clinical care for therapists who need reliable and valid psychometric assessments to identify the resources developed by patients with the aim to adapt their psychological support accordingly.

This study was therefore undertaken to determine the most suitable structure for the French version of the PTGI derived from the translation realized by Lelorain et al. [11] in early-stage breast cancer and melanoma patients. Specifically, this study addresses the structural validity and reliability of the questionnaire through confirmatory and exploratory analyses. It also assesses the concurrent validity with a coping measure and targets the association between PTGI scores and cancer location.

Material and methods

The PTGI questionnaire

The PTGI is a self-report questionnaire composed of 21 items organized into five domains: *Relating to Others* (RO, 7 items), *New Possibilities* (NP, 5 items), *Personal Strength* (PS, 4 items), *Spiritual Change* (SC, 2 items), and *Appreciation of life* (AL, 3 items) [9]. All items are scored on a 6-point Likert response format indicating the extent to which the listed changes occurred in the respondents' lives as a result of an identified trauma. Response categories range from 0 = "*I did not experience this change as a result of my crisis*" to 5 = "*I experienced this change to a very great degree as a result of my crisis*". Each domain score is computed as the sum of its item responses. A high score on a domain indicates a higher degree of reported positive changes.

In the seminal article on the PTGI, Tedeschi and Calhoun worked on a sample of students at a U.S. university who had experienced a significant negative life event in the past 5 years [9]. They performed a principal component analysis (PCA) on 34 items, followed by a varimax rotation. It produced six factors with eigenvalues greater than one. As only five factors were easily interpretable, they only retained the 21 items loading on these factors. They then performed a second PCA followed by a varimax rotation on these 21 items. It produced five factors with eigenvalues greater than one, identical to those found with 34 items. These factors are the one mentioned above (i.e., *Relating to Others*, *New Possibilities*, *Personal Strength*, *Spiritual Change*, and *Appreciation of Life*). This analysis yielded high factor loadings (0.59 to 0.85), except for two items (item 1, factor loading = 0.50: "*I changed my priorities about what is important in life*", and item 12, factor loading = 0.54: "*I am better able to accept the way things work out*"). The internal consistency of the five factors that emerged was acceptable (Cronbach's alpha coefficients between 0.67 and 0.85), as was the internal consistency of all items (Cronbach's alpha coefficient equal to 0.90). In addition, corrected item-total PTGI correlations (the correlation of each item with the total score across all remaining 20 items) ranged from 0.35 to 0.63. Pearson's correlations among factors ranged from 0.27 to 0.52, and correlations of factors with the

PTGI total score ranged from 0.62 to 0.83. Test-retest reliability was assessed over 2 months and seemed globally acceptable. Since then, the factorial structure of the PTGI has been challenged, with some studies finding one, two, three, or four factors [14–20]. Of note, some of these studies used exploratory factor analyses instead of confirmatory factor analyses (CFA), which is the recommended method for confirming a hypothesized factor structure.

Finally, Tedeschi and Calhoun did not find a significant relationship between time elapsed since the negative life event (ranging from less than 6 months to more than 4 years) and the PTGI total score [9], but inconsistent findings have been found in the literature [10]. For instance, longitudinal studies in psycho-oncology showed that PTG could rapidly occur after a cancer diagnosis and then remain stable or increase over almost 2 years [21–23]. At last, sex differences are mostly consistent among studies, with women usually reporting slightly higher PTGI scores than men [10].

French versions of the PTGI

In the French language, the PTGI has been translated twice. The first translation was proposed by Lelorain et al. [11], and the second one was realized by Cadell et al. [12]. To date, at least three French versions are currently used in France:

- Version 1: The original translation realized by Lelorain et al. [11]. In the literature, no information is available on the translation process used. We learned from the authors that they translated the questionnaire into French and that a native English-speaking professional performed a back-translation. However, we do not know whether any discrepancies were highlighted and whether they were accounted for. Of note, no content validity assessment has been performed. This French version has been adapted for cancer patients who were asked to rate items from 0 = "*Not at all*" to 5="*Totally*" to indicate whether cancer caused the change mentioned in the item.
- Version 2: A revised version that appeared in French study protocols in 2010 (following the work of Lelorain et al. [11]). This version shows slight adaptations in wording compared to the original translation (in total, changes in the wording have been made on three items over 21). These changes in wordings from version 1 are addressed in detail in supplementary material (Appendix A).
- Version 3: The translation performed by Cadell et al. [12]. To translate the PTGI from English to French, Cadell et al. used a professional translator. This French version was then independently back-translated into English by another certified translator and no differences were highlighted. The content was validated independently by a native French speaker. Of note, these authors justified their choice to re-translate the PTGI in French because they found no published version of the translation realized by Lelorain et al. [11]. Their French wording of the items are available in their manuscript [12].

Table 1 summarizes the information available on these different versions in the literature. It synthesizes the psychometric properties of each version and lists the studies where they are used as a measurement tool.

On the one hand, the psychometric properties of the two first versions (versions 1 and 2) have never been comprehensively evaluated. Indeed, the only available information regarding the original translation is that a hierarchical CFA with five first-order factors (RO, NP, PS, SC, and AL) and one second-order factor (global PTG) did not fit data from a sample composed of 307 French women who had recovered from breast cancer for at least 5 years [11]. Nonetheless, all five domains and all 21 items showed good internal consistency [11]. After further investigation, the performed CFA might have been too restrictive as no correlation between the error terms was allowed. In addition, no information on convergent and divergent validity is available. As for the version derived from the original translation of Lelorain et al. [11], no information is currently available. To the best of our knowledge, these two versions are the ones that are mainly used in France.

On the other hand, Cadell et al. validated their own French version of the PTGI on a small Canadian sample of French-speaking caregivers (combining 10 bereaved HIV/AIDS caregivers and 37 parents caring for a child with a life-limiting illness) [12]. All domains (except one, i.e., Personal strength) showed Cronbach alpha coefficients greater than 0.6 (the threshold used by the authors as providing evidence supporting reliability for exploratory studies). CFAs performed on each of the five domains indicated satisfactory evidence of convergent validity based on factor loading values. Of note, no multidimensional confirmatory factor analysis combining the five domains and their related items was performed. Dimensionality was also explored using CFAs and reported in terms of *discriminant validity*; the five domains demonstrated a high discriminant validity, indicating that they represented distinct constructs.

In brief, data on the psychometric properties of both the French translation realized by Lelorain et al. [11] (version 1) and the revised version (version 2) are too scarce to rule on the validity of these questionnaires in patients facing cancer. In addition, results emphasized by Cadell

Table 1 Overview of the available information regarding the three different French versions of the 21-item posttraumatic growth inventory

PTGI French versions	Availability of data on psychometric properties	Reported statistical analyses					Additional results	French studies involving the version
		Cronbach's alpha	Inter-item correlation	CFA + Goodness of fit	Convergent validity	Divergent validity		
Version 1: Original translation developed by Lelorain et al.	See Lelorain et al. [11] **Sample**: $N = 307$ Long term Breast cancer survivors (France)	RO: 0.85 NP: 0.86 SC: 0.83 AL: 0.81 PS: 0.79 Total PTGI (21 items): 0.93	Not available	Hierarchical CFA with 5 first-order factors (one for each domain) and one second-order factor (global PTG) The model did not fit their data well	Not avaible	Not available	No	- Three studies used this version to assess PTG in breast cancer patients [11, 24, 25] - One study used this version to validate the French translation-adaptation of the impact of cancer questionnaire version 2 (VALIOC) [26] - One study used this version to assess PTG in survivors of intimate partner violence [13]
Version 2: Revised version based on the translation of Lelorain et al.	No	Not available	Not available	Not available	Not available	Not available	No	- Three studies used this version to assess PTG in cancer survivors: ELCCA [27], VICAN [28], and EPICURE [29] - One study used this version to assess PTG in cancer patients and health care workers during Covid-19 pandemic (PAPESCO-19) [30] - One study used this version to assess PTG in renal failure patients waiting for first kidney transplantation (PreKitQol) [31]

Table 1 (continued)

PTGI French versions	Availability of data on psychometric properties	Reported statistical analyses					Additional results	French studies involving the version
		Cronbach's alpha	Inter-item correlation	CFA + Goodness of fit	Convergent validity	Divergent validity		
Version 3: Translation of Cadell et al.	See Cadell et al. [12] **Sample:** *N* = 10 Bereaved caregivers (Canada) + *N* = 37 Parents caring for a child with a life-limiting illness (Canada)	RO: 0.85 NP: 0.83 SC: 0.86 AL: 0.64 PS: 0.34 (0.77 when omitting item 12[a]) Total PTGI (21 items): 0.87	Within each domain, all items were significantly correlated, except for item 12[a]	Five unidimensional CFAs (one per domain) + One CFA on the five mean scores to the domains *Remark: No multidimensional CFA combining the five domains and the associated items was performed*	All standardized factor loadings from the five CFAs were above 0.5 All standardized factor loadings from the CFA performed on the mean scores were above 0.5 (except for AL score)	Not available	Dimensionality was explored and reported in terms of *discriminant validity*. It was assessed for each pair of domains by comparing the chi-square statistic of two CFAs: 1) A CFA where the correlation between the two factors is constrained to be equal to 1 2) A CFA where the correlation is freely estimated	- One study used this version to assess PTG after a hematopoietic stem-cell transplantation [32] - One study used this version to assess PTG in survivors of a diffuse large B-cell lymphoma [33]

PTGI Posttraumatic growth inventory, *PTG* Posttraumatic growth, *RO* Relating to others, *NP* New possibilities, *PS* Personal strength, *SC* Spiritual change, *AL* Appreciation of lie, *N* Sample size, *CFA* Confirmatory Factor Analysis

[a] Item12: "*I am better able to accept the way things work out*", Cadell French translation: "*J'accepte plus facilement la tournure que prennent les événements*"

et al. [12] may not be transposable to a French population for several reasons. First, French-speaking Canadians may not be representative of French people due to cultural differences. Besides, caregivers and patients experiencing a major health event (such as cancer) might perceive the PTG construct and PTGI questionnaire differently [15, 34]. Finally, the translation of Cadell et al. [12] is quite different from the other two versions with regard to the formulation of the items.

Patients sample

To assess the psychometric properties of the revised French version of the PTGI (version 2), we used the data collected within the ELCCA cohort (ClinicalTrials.gov Identifier: NCT02893774) [35]. ELCCA is a longitudinal study conducted in France on patients recently diagnosed with early-stage breast cancer or melanoma at the Nantes Cancerology Institute (for breast cancer patients) and the Department of Onco-Dermatology of Nantes University Hospital (for melanoma patients). It was approved by an ethical research committee ("Comité de Protection des Personnes"). The objective of ELCCA was to study socioeconomic, psychological and HRQoL changes following a breast cancer or a melanoma within the months and years following the cancer diagnosis. Breast cancer and melanoma were chosen as they have a similarly good prognosis when detected early. Nevertheless, health care is very different for these two cancer locations. On the one hand, breast cancer treatments are generally invasive with major surgery and therapies such as radio-, chemo- and hormonal therapies. On the other hand, melanoma patients experience a minor surgery possibly followed by immunotherapy [35]. Hence, the different nature of the treatments could interfere with the perceived severity of the disease, and by extension, with the changes that patients may experience.

After the diagnosis, patients were informed about the study by an oncologist and were invited to sign an informed consent agreement. Participants were asked to complete a series of questionnaires at different measurement occasions: 1, 6, 12, 24, 48, and 60 months post-diagnosis. We chose to analyze the data from the 4th measurement occasion (i.e. 2 years post-diagnosis) to allow PTG to take place and allow patients sufficient time to report it [10]. In addition to questionnaires, patients had to report socioeconomic and clinical information. They completed the booklet of questionnaires at home or during a follow-up visit.

Measures

At each measurement occasion, patients were asked to complete a series of questionnaires that included the PTGI (version 2: revised French version based on Lelorain et al. translation) [9, 11] and the Brief COPE [36, 37]. Of note, the Brief COPE is an abbreviated version of the COPE [38]. It contains 14 subscales (composed of two 4-point Likert items each) aiming to assess the coping strategies used by the patients (active coping, planning, using instrumental support, using emotional support, venting, behavioral disengagement, self-distraction, self-blame, positive reframing, humor, denial, acceptance, religion, and substance use). Subscale scores are computed as the sum of the item responses, with a higher score indicating a higher use of a given strategy to deal with stressful life events.

Statistical analysis

Confirmatory analysis

To determine whether the initial five-factor structure of the PTGI was suitable for the French version used in the ELCCA study, we conducted a confirmatory analysis aiming to inform its structural validity and reliability.

First, we conducted a confirmatory factor analysis (CFA) with oblique factors based on the initial five-factor structure of the PTGI using maximum likelihood estimation [39, 40]. The chi-square statistic used to assess the goodness of fit was corrected using Satorra-Bentler adjustment to obtain results robust to non-normality (item responses being in an ordinal format with a 6-point response scale). Good (or acceptable) fit was indicated by the following criteria: RMSEA ≤ 0.05 (≤ 0.08), CFI ≥ 0.95 (≥ 0.90) and SRMR ≤ 0.05 (≤ 0.10). A hierarchical CFA was also conducted to explore a second-order factor structure representing global PTG.

We assessed the item-level convergent validity of each domain to ensure that the items correlated well with their hypothesized domain. It was evaluated by examining the Spearman's item-rest correlations (i.e., the Spearman's correlation coefficient between the item score and the domain score computed without the item). Item-level convergent validity was considered good (or acceptable) when 100% (respectively 95%) of the items from a given domain had an item-rest correlation greater than 0.4. Item-level divergent validity was also evaluated to ensure that items were more correlated with their own (hypothesized) domain than with the other domains of the scale. It was assessed by comparing each Spearman's item-rest correlation coefficient (computed between a given item and its hypothesized domain) with Spearman's correlation coefficients computed between the studied item and the other domains. The item-level divergent validity was considered good (or acceptable) when 100% (respectively 95%) of the items of the questionnaire were more correlated with their hypothesized domain than with the other domains. Correlations between PTGI domains were also examined using Spearman's correlation coefficients. They

were expected to be moderately to highly correlated with one another as found in previous validation studies of the PTGI in cancer patients [41, 42].

Internal consistency of the PTGI domains was assessed using the Cronbach's alpha coefficient α [43]. Domains were considered reliable if $\alpha > 0.7$. We also used backward Cronbach alpha curves to determine whether domains were unidimensional or not [44]. These curves were obtained for each domain by: (1) Computing the Cronbach's alpha coefficient over all items of the considered domain, (2) Removing items one by one until there are only two items left. At every successive step, the removed item is the one that left the remaining set of items with the maximum Cronbach's alpha. If the set of items is not unidimensional, increases in the Cronbach's alpha coefficient will be observed after item removal. In addition, the Loevinger's coefficients (H and H_i) in relation to the domains and to the items, respectively, were used to evaluate the homogeneity of the domains ($H > 0.3$ indicating a high degree of homogeneity) and to determine whether a given item i was consistent with its domain (verified if $H_i > 0.3$) [45, 46].

Of note, personal mean score imputations were realized in case of partial missing data in one or several domains. Within each of the five domains, missing data were imputed by the personal mean score if the number of non-missing items was higher than half the number of items comprising that domain.

Exploratory analysis

In case of unsatisfying results, we planned to conduct a clustering of variables around latent components (CLV) analysis to identify the most optimal structure for the French version of the PTGI.

CLV is an exploratory analysis which has been developed by Vigneau and Qannari [47], and aims to identify unidimensional and disjointed sets of items. To prevent the imputation performed during the confirmatory analysis from possibly favoring the emergence of the five initial factors,[1] we planned to:

- Step 1: Conduct the CLV analysis on the raw data (not imputed by the personal mean score)
- Step 2: Impute missing data using the personal mean score based on the structure evidenced by the CLV analysis during step 1
- Step 3: Ensure that the evidenced structure remains stable by re-running the CLV analysis on this newly imputed data set

All analyses already described for the confirmatory analysis were performed to assess the structural validity and reliability of a new structure potentially evidenced by the CLV analysis (i.e. CFA, item-level convergent and divergent validity, and reliability).

Concurrent validity and comparison of PTGI scores according to cancer location

We explored the concurrent validity and compared the PTGI scores according to cancer location based on the most optimal structure for the French version of the PTGI studied in this manuscript (structure evidenced by either the confirmatory factor analysis or the exploratory analysis).

Concurrent validity was evaluated by assessing the plausibility of a priori assumptions about patterns of association between the PTGI scores and the Brief COPE scores. Tedeschi and Calhoun asserted that PTG was both a process and an outcome [48]. Specifically, PTG could be a strategy for coping, managing and surviving trauma just after it occurs (i.e., a process), and then turn into an outcome at a later time, as positive changes can be expressed after a challenging experience [10]. Hence, we expected positive correlations between PTGI scores and the following domains of the Brief COPE: active coping, planning, using instrumental support, using emotional support, venting, positive reframing, humor, and acceptance. The hypothesized domain comprising the item 5 and 18 (related to a Spiritual Change according to Tedeschi and Calhoun's theoretical rationale [9]) was also expected to be strongly correlated with religious coping.

The association between the PTGI scores and the cancer location were examined using Mann-Whitney tests. PTGI scores were expected to be higher for breast cancer patients, as breast cancer patients may experience more social support than melanoma patients [35, 49] and since previous research suggested that experience sharing and social support are associated with PTG [50]. To withdraw the gender effect in the comparison of melanoma (both women and men) and breast cancer patients (only women), we restricted this score comparison to the data collected in women.

All analyses were performed with Stata 16 (*Stata Statistical Software: Release 16*. College Station, TX: StataCorp LLC). The Stata module used to perform CLV is available from the Statistical Software Components archive [51]. The French version of the PTGI evaluated in this manuscript (i.e., the version 2 that derived from Lelorain et al. translation [11]) is available in the supplementary materials (see the Appendix B).

[1] During the confirmatory analysis, missing data were imputed by the personal mean score based on the initial five-factor structure.

Results

Sample characteristics

At the 4th measurement occasion (i.e., 2 years after their cancer diagnosis), 380 participants sent back the PTGI questionnaire. Among them, one patient with breast cancer only responded to one item, she was excluded from the analysis. Of the 379 remaining participants, 299 (79%) had breast cancer and 80 (21%) had melanoma. They were aged between 21 and 73 years, with an average age of 54.8 (SD = 9.1 years). All participants with breast cancer were women. Among melanoma patients, 32 were women (40%) and 48 were men (60%). Most of them lived in couple (81%). A total of 361 out of 379 (95%) individuals had a complete PTGI (i.e., no missing data in their responses). Item-level missing data rates ranged from 0.3 to 1.8% (this latter rate being reached by item 18: "*I have a stronger religious faith*").

Confirmatory analysis

Confirmatory factor analysis

The confirmatory factor analysis based on the initial five-factor structure of the PTGI (with oblique factors) indicated an acceptable fit suggested by the goodness of fit criteria (with Satorra-Bentler correction): $\chi^2(179, n = 369) = 507.4$, $p < 0.001$, $\chi^2/df = 2.8$, RMSEA = 0.071, CFI = 0.909, SRMR = 0.048. All standardized factor loadings were greater than 0.55; they are given in Table 2. High covariances ranging from 0.76 to 0.90 were observed among four of the five factors: *Relating to others*, *New possibilities*, *Personal strength* and *Appreciation of Life*. When the *Spiritual change* factor was involved, covariance did not exceed 0.43. Of note, adding a second-order factor (i.e., global PTG) through a hierarchical CFA did not improve the model fit: $\chi^2(184, n = 369) = 521.5$, $p < 0.001$, $\chi^2/df = 2.8$, RMSEA = 0.071, CFI = 0.906, SRMR = 0.050. In this hierarchical CFA, the factor loading associated with the *Spirituality change* factor was low (0.42). Together with the moderate covariances observed between the *Spiritual change* factor and the other factors within the oblique CFA, these results indicate that the PTGI should be considered a multidimensional scale. Hence, we did not consider a total score computed over the 21 items in the following. Finally, adding covariance between the error terms of two items from the same domain (based on large modification indices) did not really improve model fit.

Item-level convergent-divergent validity

All items had a Spearman's item-rest correlation with their hypothesized domain greater than 0.4, confirming the convergent validity. However, criterion for divergent validity was not met; indeed only 17 items over the 21 (i.e., 81%) were more correlated with their hypothesized domain than with the other domains. Specifically, all items from the *Relating to others*, *New possibilities,* and *Spiritual Change* domains were more strongly correlated with their hypothesized domain than with other ones. However, some items from the *Personal strength* and *Appreciation of Life* domains were divergent:

- Item 12 (*Personal Strength* domain): "*I am better able to accept the way things work out*" was more strongly correlated with the *New possibilities* and *Appreciation of life* domains (Spearman's correlation coefficients were 0.62 and 0.61, respectively) than with the *Personal strength* domain (Spearman's item-rest correlation was 0.60). Nevertheless, the difference between the correlations remained small.
- Item 1 (*Appreciation of life*): "*I changed my priorities about what is important in life*" had a Spearman's correlation coefficient of 0.59 with the *New possibilities* domain while its Spearman's item-rest correlation with the *Appreciation of life domain* was 0.47.
- Item 13 (*Appreciation of life*): "*I can better appreciate each day*" correlated more with two other domains: *New possibilities* and *Personal strength* (correlation coefficients were respectively 0.61 and 0.64) than with *Appreciation of life* (item-rest correlation of 0.55).

Correlation between the five PTGI domains

Spearman's correlation coefficients between the *Relating to others*, *New possibilities*, *Personal strength* and *Appreciation of life* domains were high (ranging from 0.62 to 0.72). Spearman's correlation coefficients between the *Spiritual change* domain and the other domains were small to moderate (between 0.25 and 0.36).

Reliability

All domains showed good internal consistency with Cronbach's alpha coefficients ranging from 0.75 to 0.88. Loevinger coefficients related to the domains (H) varied between 0.54 and 0.80, indicating acceptable homogeneity for all domains. At the item level, all items exhibited Loevinger coefficients greater than 0.3. Hence, all items seemed to be consistent within their domain. All coefficients regarding reliability are given in Fig. 1 alongside the backward Cronbach alpha curves. Of note, a dimensionality issue appeared within the *Appreciation of Life* domain: the associated backward Cronbach alpha curve increased when removing item 1. It indicated that this domain might not be unidimensional.

Table 2 Standardized factor loadings of the 21 items of the posttraumatic growth inventory obtained after the oblique confirmatory factor analysis based on the initial five-factor structure

	Standardized factor loadings for each factor				
PTGI items	**RO**	**NP**	**PS**	**SC**	**AL**
Relating to others (RO)					
6. I more clearly see that I can count on people in times of trouble *Je vois mieux que je peux compter sur les autres en cas de problème*	0.56	–	–	–	–
8. I have a greater sense of closeness with others *Je me sens plus proche des autres*	0.81	–	–	–	–
9. I have a greater willingness to express my emotions *Je suis plus enclin(e) à exprimer mes émotions*	0.76	–	–	–	–
15. I have greater compassion for others *J'ai plus de compassion pour les autres*	0.62	–	–	–	–
16. I put more effort into my relationships *J'investis plus mes relations aux autres*	0.81	–	–	–	–
20. I learned a great deal about how wonderful people are *Je vois plus le bon côté des gens*	0.77	–	–	–	–
21. I better accept needing others *J'accepte mieux le fait d'avoir besoin des autres*	0.71	–	–	–	–
New possibilities (NP)					
3. I developed new interests *Je me suis intéressé(e) à de nouvelles choses*	–	0.77	–	–	–
7. I established a new path for my life *J'ai donné une nouvelle direction à ma vie*	–	0.77	–	–	–
11. I'm able to do better things with my life *Je fais de ma vie quelque chose de meilleur*	–	0.83	–	–	–
14. New opportunities are available which wouldn't have been otherwise *De nouvelles opportunités sont apparues*	–	0.67	–	–	–
17. I'm more likely to try to change things which need changing *J'essaie davantage de changer les choses qui ont besoin d'être changées*	–	0.72	–	–	–
Personal strength (PS)					
4. I have a greater feeling of self-reliance *J'ai acquis plus confiance en moi*	–	–	0.75	–	–
10. I know better that I can handle difficulties *Je suis davantage capable de gérer des situations difficiles*	–	–	0.77	–	–
12. I am better able to accept the way things work out *J'accepte mieux la façon dont les choses se passent*	–	–	0.78	–	–
19. I discovered that I'm stronger than I thought I was *J'ai découvert que je suis plus fort(e) que ce que je pensais*	–	–	0.68	–	–
Spiritual change (SC)					
5. I have a better understanding of spiritual matters *J'ai développé une certaine spiritualité*	–	–	–	0.94	–
18. I have a stronger religious faith *J'ai une foi religieuse plus grande*	–	–	–	0.83	–
Appreciation of life (AL)					
1. I changed my priorities about what is important in life *J'ai changé de priorités dans la vie*	–	–	–	–	0.59
2. I have a greater appreciation for the value of my own life *J'apprécie plus ma vie à sa vraie valeur*	–	–	–	–	0.79
13. I can better appreciate each day *J'apprécie plus amplement chaque jour de ma vie*	–	–	–	–	0.83

The wording of the items in the French version of the PTGI that we studied (version 2) are shown in italics

Summary

Based on statistical results, the studied French version of the PTGI showed mixed psychometric properties. Indeed, the fit of the CFA was acceptable and the convergent validity was good. However, the *Appreciation of Life* domain seemed problematic: two items over three

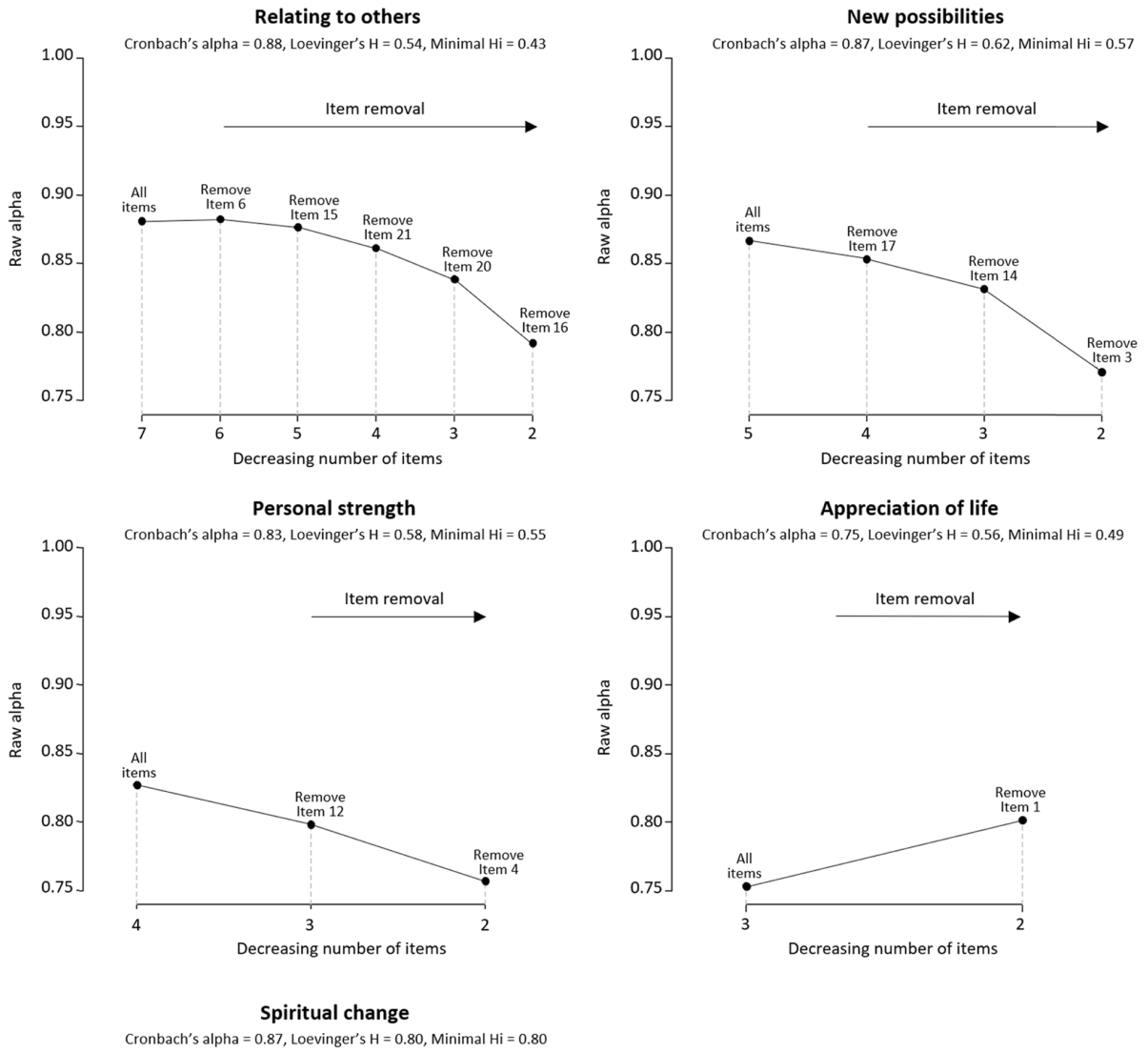


Fig. 1 Cronbach's and Loevinger's coefficients for the five initial domains and backward Cronbach alpha curves for the four domains composed of three items or more

Figure notes: - Minimal Hi: Minimal Loevinger's coefficient at the item level over all items of the domain. - The Spiritual Change domain is composed of two items, hence no backward Cronbach alpha curve can be drawn

correlated more with other domains, and the alpha curve emphasized a dimensionality issue.

We also noticed several translational issues. For instance, in the wording of French items, the notion of ability has completely disappeared for some items (e.g., items 11 and 12). In addition, the wording of the French response categories is different. Indeed, instead of describing a degree of change (as in the initial version of Tedeschi and Calhoun [9]), Lelorain et al. chose to use the following response categories: 0 = "*Not at all*", 1 = "*Very few*", 2 = "*Few*", 3 = "*A little*", 4 = "*A lot*", and 5 = "*Totally*" [11]. Yet, in the French language, the response categories 1 and 2 are hardly differentiable, which might confuse the respondents.

To overcome these issues, we grouped the response categories 1 and 2, and conducted an exploratory study to identify the most optimal structure for this French version of the PTGI in terms of validity and reliability.

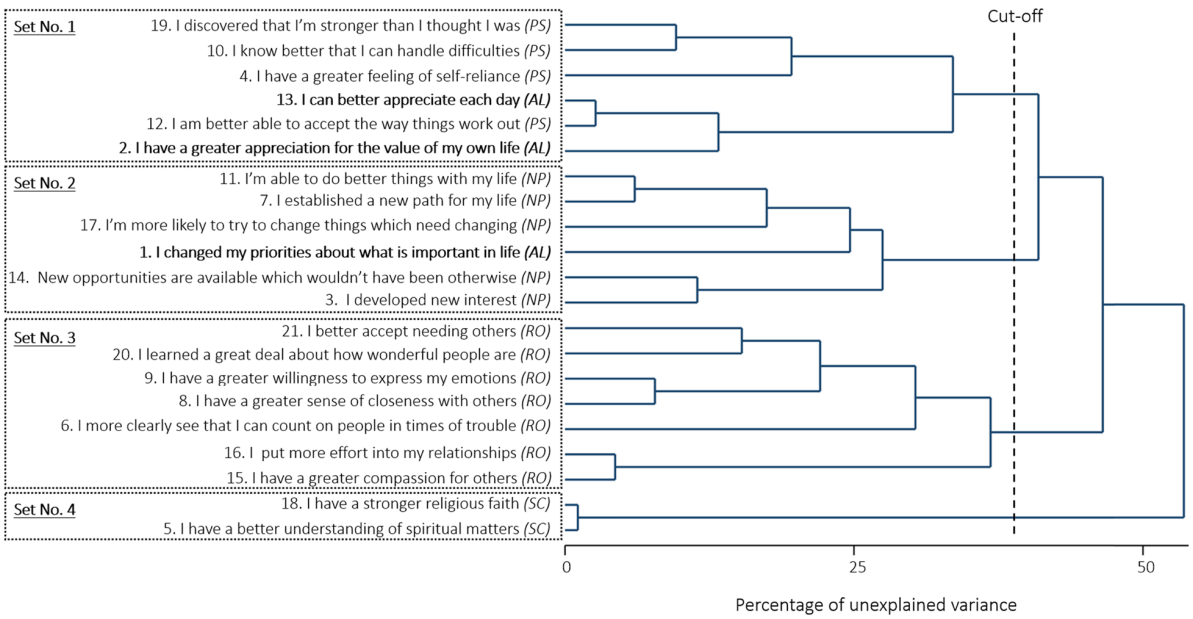


Fig. 2 Dendrogram of the Clustering around Latent Variable analysis. Initial domains are indicated in brackets. New unidimensional sets of items are emphasized by boxes

Figure notes: PS: Personal strength, AL: Appreciation of life, NP: New possibilities, RO: Relating to others, SC: Spiritual change

Exploratory analysis

Clustering of variables around latent components

The CLV analysis was performed on the raw data (not imputed, $n = 361$) where response categories 1 and 2 were grouped beforehand. The analysis led to four unidimensional sets of items that were close to the initial five domains of the PTGI (the dendrogram is given in Fig. 2):

- **Set of items No. 1 [*Personal capacities*]:** All items from the *Personal strength* domain plus two items from the *Appreciation of Life* domain (items 2 "*I have a greater appreciation for the value of my own life*" and 13 "*I can better appreciate each day*")
- **Set of items No. 2 [*New life direction*]:** All items from the *New possibilities* domain plus one item from the *Appreciation of life* domain (item 1 "*I changed my priorities about what is important in life*")
- **Set of items No. 3 [*Relating to others*]:** All items from the *Relating to others* domain
- **Set of items No. 4 [*Spiritual change*]:** All items from the *Spiritual Change* domain

Hence, four of the five initial domains were recovered. The fifth domain (i.e., *Appreciation of life*) was not. The three items composing it have been separated; two of them were grouped with the items from the *Personal strength* domain, and the other one was grouped with the items from the *New possibilities* domain. We labelled the sets of items No. 1 and No. 2 *Personal capacities* and *New life direction*, respectively, as these labels better reflected the items composing these new domains and the groupings that have been made. In this newly evidenced four-factor structure, the sum score goes from 0 to 24 for the sets of items No. 1 and No. 2 (*Personal capacities* and *New life direction, 6 items each*), from 0 to 28 for the set No. 3 (*Relating to others, 7 items*), and from 0 to 8 for the set No. 4 (*Spiritual change, 2 items*). Of note, the dendrogram remained the same when re-running the CLV analysis after the missing data imputation by the personal mean score based on this new four-domain structure.

Confirmatory factor analysis

The confirmatory factor analysis based on the newly evidenced four-factor structure indicated an acceptable fit suggested by the goodness of fit criteria (with Satorra-Bentler correction): $\chi^2(183, n= 371) = 524.0$, $p < 0.001$, $\chi^2/df = 2.9$, RMSEA = 0.071, CFI = 0.908, SRMR = 0.048. All standardized factor loadings were above 0.55, they are given in the supplementary materials (Appendix C). Strong covariances among *Relating to others*, *New*

life direction and *Personal capacities* factors were again observed (they were all above 0.80). Covariances involving the *Spiritual change* factor were lower and did not exceed 0.41.

Item-level convergent-divergent validity

All items had an item-rest correlation with their hypothesized domain greater than 0.4, confirming the convergent validity. In addition, all items exhibited larger correlation coefficients with their domain than with the other ones (except item 11 "*I'm able to do better things with my life*", which correlated slightly more with the *Personal capacities* domain than with the *New life direction* domain: Spearman's correlation coefficients were 0.73 versus 0.69). This result indicated an acceptable divergent validity.

Correlation between the four new PTGI domains

Spearman's correlation coefficients among the *Relating to others*, *New life direction*, and *Personal capacities* domains were high (ranging from 0.68 to 0.72). Spearman's correlation coefficients between the *Spiritual change* domain and the other domains were moderate (between 0.30 and 0.33).

Reliability

All domains showed good internal consistency and homogeneity as Cronbach alphas were all greater or equal to 0.85 and Loevinger coefficients were greater than 0.4 (values for each domain are given in Fig. 3). In addition, backward Cronbach alpha curves did not evidence dimensionality issues since all curves decrease with item withdrawal (see the Fig. 3).

Summary

This new four-factor structure shows better psychometric properties than the five-factor structure. Hence, the rest of the analyses (i.e., concurrent validity and comparison of PTGI scores according to cancer location) will be performed on this structure comprising four domains.

Concurrent validity

Spearman's correlation coefficients between the new PTGI scores and the Brief Cope scores are given in Table 3, part (a). As expected, we observed positive correlations between the new PTGI domains *Relating to others*, *New life direction* and *Personal capacities* and the following domains of the Brief COPE: active coping, planning, using instrumental support, using emotional support, venting, positive reframing, acceptance, and humor. Of note, correlations between acceptance, emotional support and the three above-mentioned domains of the PTGI were nonetheless small. The strong link between religious coping and the *Spiritual change* domain was also confirmed, but religious coping was actually positively correlated with all the PTGI domains.

PTGI scores according to the cancer location

As expected, breast cancer women showed significantly higher PTGI scores than melanoma women with a significance level of 5%, except for the *Relating to others* domain for which no significant difference was found. Description of the PTGI scores according to the cancer location among women is available in Table 3, part (b).

Discussion

Main results

In this current study, we first examined the psychometric properties of the French version of the PTGI in early-stage breast cancer and melanoma patients based on the initial five-factor structure [9]. This questionnaire showed mixed psychometric properties. Indeed, while the fit of the CFA was acceptable and the convergent validity was good, the *Appreciation of Life* domain seemed problematic. In addition, we noticed several translational issues in the items wording and in the response categories. Hence, we decided to group two response categories that were hardly differentiable in French language and searched for the optimal structure for this French version of the PTGI in terms of validity and reliability.

Based on an exploratory factor analysis, we evidenced a four-factor structure close to the initial five-factor structure. Indeed, four of the five initial domains were recovered, and the items from the unrecovered domain, i.e. *Appreciation of life*, have been separated; two of them were grouped with the items from the *Personal strength* domain, and the other one was grouped with the items from the *New possibilities* domain. On the one hand, item 1 ("*I changed my priorities about what is important in life*") was grouped with the items from the *New possibilities* domains. It made sense as patients probably perceived this item as dealing with the changes regarding their life orientation (e.g., things they want or no longer want to dedicate time to, things they want to change, and paths they want to follow) making this item close to items from the *New possibilities* domain. We labelled this new set of items *New life direction*, as it reflects the novelty and being active (as opposed to passive) regarding life orientation. On the other hand, items 2 ("*I have a greater appreciation for the value of my own life*") and 13 ("*I can better appreciate each day*") were grouped with the items from the *Personal Strength* domain. More

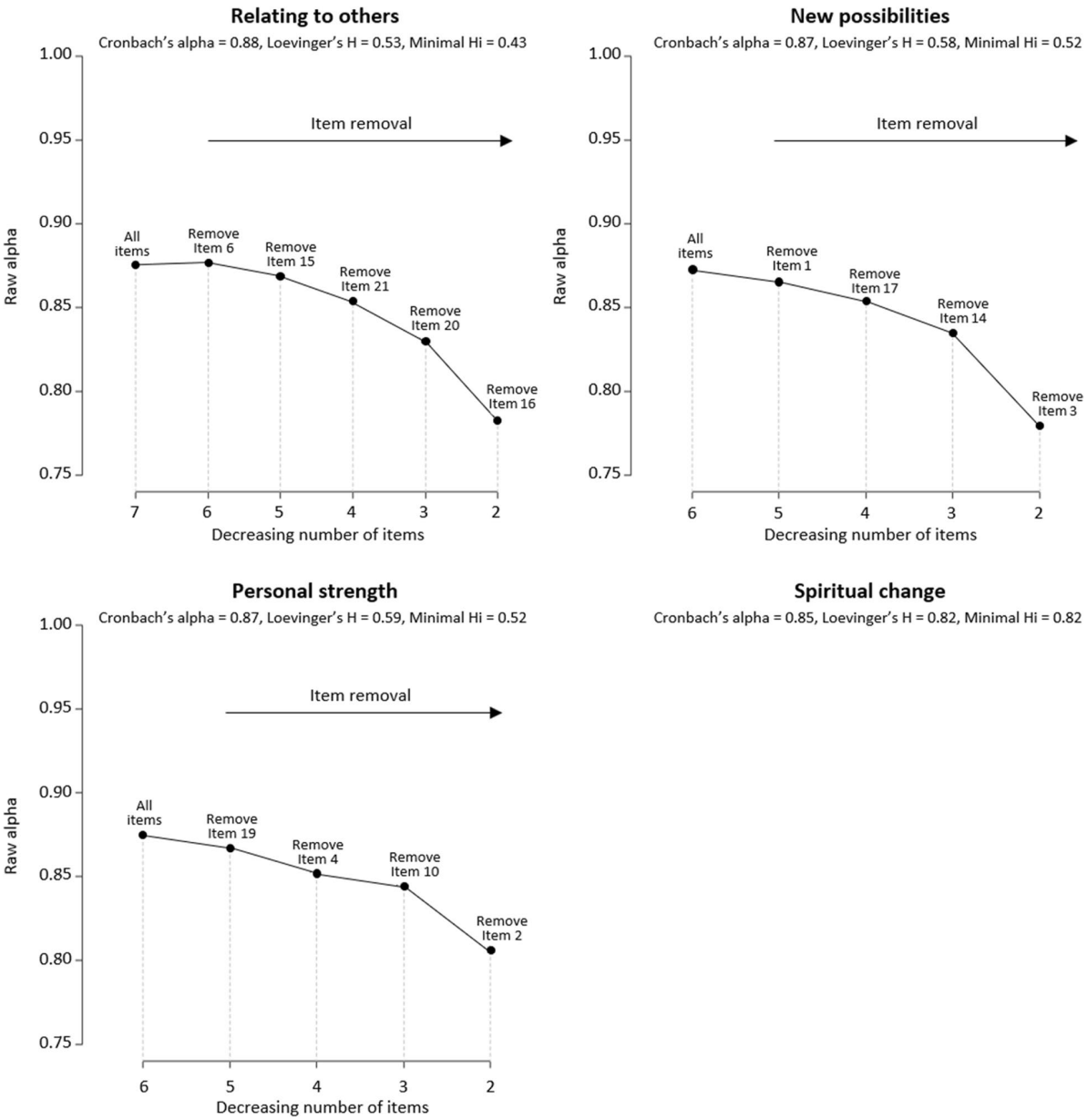


Fig. 3 Cronbach's and Loevinger's coefficients of the four new domains evidenced by the Clustering around Latent Variable analysis and backward Cronbach alpha curves for the three domains composed of three items or more

Figure notes: Minimal Hi: Minimal Loevinger coefficient at the item level over all items of the domain. The Spiritual Change domain is composed of two items, no backward Cronbach alpha curve can be drawn

precisely, we can observe in the dendrogram of the CLV analysis (Fig. 2) that these items are rapidly grouped with item 12 ("*I am better able to accept the way things work out*"). This merging might be explained by the strong correlation between items 2, 12, and 13 (ranging from 0.6 to 0.7). A possible explanation for these associations is that acceptance (targeted by item 12) has been shown to promote well-being (targeted by items 2 and 13) as reported in papers dealing with Acceptance and Commitment therapy [52–54]. We labelled this new set of items *Personal capacities*, as the confrontation with trauma

Table 3 (a) Correlations between the new PTGI scores and the Brief Cope scores and (b) associations between the new PTGI scores and the cancer location

	Relating to others	New life direction	Personal capacities	Spiritual Change
(a) Correlation with Brief Cope scores (Spearman's correlation coefficients r)				
Positive reframing	0.37*	0.40*	0.46*	0.25
Active coping	0.35*	0.41*	0.39*	0.15
Planning	0.34*	0.38*	0.30*	0.19
Humor	0.35*	0.32*	0.36*	0.03
Acceptance	0.22	0.21	0.31*	0.08
Self-distraction	0.17	0.24	0.15	0.08
Venting	0.37*	0.36*	0.28	0.11
Instrumental support	0.36*	0.35*	0.22	0.15
Emotional support	0.23	0.24	0.11	0.13
Religion	0.31*	0.28	0.22	0.75*
Self-blame	0.18	0.20	0.02	0.18
Denial	0.06	0.03	0.02	0.07
Behavioral disengagement	−0.07	−0.10	−0.13	0.01
Substance use	− 0.08	− 0.04	−0.09	− 0.04
(b) Association with cancer location (median of the sum scores for each domain and IQR)				
Women with breast cancer ($n = 299$)	14.0 [IQR = 8.0]	10.0 [IQR = 7.0]	13.0 [IQR = 6.0]	1.0 [IQR = 3.0]
Women with melanoma ($n = 32$)	14.0 [IQR = 8.5]	8.5 [IQR = 7.0]	10.5 [IQR = 7.5]	0.0 [IQR = 1.5]
p-value [a]	0.321	0.030	0.046	0.034

IQR Interquartile range, *n* sample size

*: $|r| \geq 0.3$ and *p*-value <0.05

[a] Mann-Whitney tests

resulted in the development of emotion regulation strategies.

Regarding the concurrent validity, PTGI scores to the *Relating to others*, *New life direction* and *Personal capacities* domains (based on the new four-factor structure) were positively correlated with positive and emotional copings, that is, positive reframing, active coping, planning, humor, acceptance, venting, using instrumental support and religion. These findings are consistent with our expectations and with literature focusing on cancer patients [11, 21, 55–57]. The strong positive association between religious coping and the *Spiritual change* domain was also confirmed, but religious coping was actually positively correlated with all the PTGI domains. This result is also consistent with literature [58–60]. Besides, except for the correlation between religious coping and *Spiritual change*, the correlations between PTGI and Brief COPE scores were moderate in absolute values. This result is in line with the debate about whether PTG is a coping strategy or not [10].

Finally, when compared with melanoma patients, breast cancer patients showed higher scores for all domains of the PTGI except one (i.e., *Relating to others*), for which no significant difference was evidenced. These results are in line with our expectations. A possible explanation is that breast cancer patients might perceive their disease as more severe than melanoma patients. Indeed, Bourdon et al. found in a qualitative interview-based study that melanoma may be trivialized because it is asymptomatic in its early stages [61]. In addition, research among cancer patients showed that people who perceive their cancer as a traumatic or highly stressful event are more likely to report PTG [62–64].

Of note, we could not compare our results with those obtained by Lelorain et al. [11] or Cadell et al. [12] as the retained structure of the questionnaire was not the same and because the data they provided did not match the analysis performed in our study.

Limitations and perspectives

Our study has several limitations that can be mentioned. First, inclusion criteria targeted only early-stage cancer patients from two hospital units (one per cancer site) data were collected 2 years after the cancer diagnosis for all patients. This may limit the generalizability of our results. Further validation studies across different cancer diagnoses, with various times since the onset of the disease, and in other French-speaking countries are needed to assess

the psychometric stability of the French version of the PTGI (i.e., reproducibility under conditions of limited change). In addition, during Study 2, we performed the CLV analysis and the CFA on the same sample. A CFA using another independent sample should be carried out in the future to investigate structure stability. Finally, the responsiveness of the PTGI to changes throughout remission has not been evaluated, and the ELCCA study design did not allow us to assess the test-retest reliability (as time-laps between two measurement occasions were large).

Regarding the translation into French, the original translation of Lelorain et al. [11] may not have been realized by a professional translator, and although a back translation has been performed, we do not know if discrepancies were evidenced and whether they were accounted for. When we compared the items wording of the French and English versions, we noticed two main discrepancies. First, in the French version, item 5 (which initially tackled the better understanding of spiritual matters in the English version) asks individuals whether they have experienced spiritual growth. Yet, having a better understanding of spirituality refers to spiritual changes that do not necessarily mean that patients grow in spirituality. Indeed, change in spirituality can be either characterized by growth or decline [65, 66]. Second, the reference to *ability* (e.g., *being able to* do something) disappeared in the French wording of some items. For instance, the French translation of item 12 could be back-translated as: "*I better accept the way things work out*", whereas the English version was "*I am better able to accept the way things* work *out*". On the one hand, the wording "*I am better able to accept*" may suggest that acceptance could depend on the situation the patient is facing; it provides context-related information. On the other hand, "*I better accept*" is a broader statement that could apply to all situations encountered; it gives more general information. Therefore, due to the disappearance of the reference to ability, professionals using PTGI for psychological support may miss context-related information, whereas this is often the information sought in psychotherapy [67]. Compared to Lelorain et al. [11] translation, the PTGI version we studied presents major changes in wording for two items. First, the second part of item 14 ("*New opportunities are available, which wouldn't have been otherwise*"), disappeared in the version we studied. In addition, item 20 (English version: "*I learned a great deal about how wonderful people are*") has been entirely reformulated in the studied version and could be back-translated as "*I see more the good side of people*". If the notion of goodness in people can be found in both wording, it is much more prominent in the English version. Assessing the impact of these differences in wording could be interesting, but it would require the collection of data on the original translation of Lelorain et al. [11] in the target population. Finally, the wording of the French response categories is different from the original wording of Tedeschi and Calhoun [9] . Indeed, instead of describing a degree of change, the response categories have been translated by Lelorain et al. as follows: 0 = "*Not at all*", 1="*Very few*", 2="*Few*", 3="*A little*", 4="*A lot*", and 5="*Totally*" [11]. As already mentioned, we chose to group response categories 1 and 2 as their French wordings were very close and might confuse respondents. As some discrepancies can be found between the French versions and the original version of Tedeschi and Calhoun [9] (at both the item and response categories levels), we would advise not comparing French scores with non-French speaking countries. Of note, the possibility of such shifts in the items meaning across translations has already been pointed out by Tedeschi et al. [10].

Finally, to avoid the co-existence of several versions, we would also advise publishing the translated questionnaire alongside the final report and/or validation manuscript.

Conclusion

This study draws a more comprehensive assessment of the psychometric properties of one of the most widely used PTGI French translations in cancer patients. It highlights that the process of translation and cross-cultural validation is crucial and must follow a rigorous methodology to obtain a reliable and valid assessment tool, valuable in both research and psychotherapy. Indeed, the PTGI allows to capture the positive change following a traumatic situation such as a cancer diagnosis and reflect how much an individual has positively re-assessed his/her life values and worldview [68]. This instrument allows relying on a measure of positive change in addition to the anxiety and depression measures and provides objective feedback on psychological support effectiveness. Given the results, we recommend that French professionals in psycho-oncology use the revised translation of Lelorain et al. [11] with five response categories instead of six (see the Appendix D of the supplementary materials) and avoid using the original five scores of the PTGI but rather the four scores proposed in our manuscript.

Abbreviations

AL: Appreciation of life; CFA: Confirmatory factor analysis; CFI: Comparative fit index; CLV: Clustering of variables around latent components; HIV/AIDS: Human immunodeficiency virus infection and acquired immunodeficiency syndrome; HRQoL: Health-related quality of life; NP: New possibilities; PCA: Principal component analysis; PS: Personal strength; PTG: Posttraumatic growth; PTGI: Posttraumatic growth inventory; RMSEA: Root mean square error of approximation; RO: Relating to others; SC: Spiritual change; SRMR: Standardized root mean squared residual.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-022-01722-6.

Additional file 1.

Acknowledgments
The authors would like to thank Pr. Mario Campone (the investigator of the ELCCA cohort), and Dr. François Dravet and Pr. Gaëlle Quereux who included the patients. The authors also sincerely thank the patients who participated in the study, the "Qualité de vie et cancer" platform, and Sophie Lelorain for the information she gave us about her translation.

Study record detail
ClinicalTrials.gov, identifier: NCT02893774.

Authors' contributions
YD analyzed the data and drafted the manuscript. MYB, VS, MAB, and JBH revised substantially the article. VS and MBourdon participated in acquiring the financial support for the project leading to this publication. All authors read and approved the final manuscript.

Availability of data and materials
Research data are not shared. The French version of the posttraumatic growth inventory evaluated in the manuscript is available in the supplementary materials (Appendix B) alongside the final version of the questionnaire (Appendix D). Analysis code for this study is available by emailing the corresponding author.
Funding: ELCCA was financially supported by the "Ligue Nationale contre le Cancer" and the "Institut pour la Recherche en Santé Publique". Y. Dubuy received a national grant from the French Ministry of Higher Education, Research and Innovation, and M. Bourdon was funded by the SIRIC ILIAD INCA-DGOS-Inserm 12558 grant.

Declarations

Ethics approval and consent to participate
The ELCCA cohort has been approved by an ethical research committee ("Comité de protection des personnes") prior to being carried out and all participants signed an informed consent agreement. All methods were performed in accordance with the relevant guidelines and recommendations, among which the declarations of Helsinki and the French and international guidelines for good clinical practice.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Nantes Université, Université de Tours, INSERM, MethodS in Patients-Centered Outcomes and HEalth Research, SPHERE, F-44000 Nantes, France. [2]Nantes Université, CHU Nantes, Methodology and Biostatistics Unit, F-44000 Nantes, France. [3]Integrated Center for Oncology (ICO, Institut de Cancérologie de l'Ouest) - Nantes, Angers, France. [4]Nantes Université, CHU Nantes, Public Health Department, F-44000 Nantes, France.

Received: 14 March 2022 Accepted: 2 September 2022
Published online: 24 September 2022

References

1. Elliott J, Fallows A, Staetsky L, Smith PWF, Foster CL, Maher EJ, et al. The health and well-being of cancer survivors in the UK: findings from a population-based survey. Br J Cancer. 2011;105(Suppl 1):S11–20.
2. Härtl K, Schennach R, Müller M, Engel J, Reinecker H, Sommer H, et al. Quality of life, anxiety, and oncological factors: a follow-up study of breast Cancer patients. Psychosomatics. 2010;51(2):112–23.
3. INCa. La vie deux ans après un diagnostic de cancer - De l'annonce à l'après-cancer [Internet report in French]. 2014 [cited 2021 Oct 11]. Available from: https://www.e-cancer.fr/Expertises-et-publications/Catalogue-des-publications/La-vie-deux-ans-apres-un-diagnostic-de-cancer-De-l-annonce-a-l-apres-cancer
4. Jefford M, Ward AC, Lisy K, Lacey K, Emery JD, Glaser AW, et al. Patient-reported outcomes in cancer survivors: a population-wide cross-sectional study. Support Care Cancer. 2017;25(10):3171–9.
5. Maass SWMC, Roorda C, Berendsen AJ, Verhaak PFM, de Bock GH. The prevalence of long-term symptoms of depression and anxiety after breast cancer treatment: a systematic review. Maturitas. 2015;82(1):100–8.
6. INCa. La vie cinq ans après un diagnostic de cancer [Internet report in French]. 2018 [cited 2021 Oct 6]. Available from: https://www.e-cancer.fr/Expertises-et-publications/Catalogue-des-publications/La-vie-cinq-ans-apres-un-diagnostic-de-cancer-Rapport
7. Sears SR, Stanton AL, Danoff-Burg S. The yellow brick road and the emerald city: benefit finding, positive reappraisal coping and posttraumatic growth in women with early-stage breast cancer. Health Psychol Off J Div Health Psychol Am Psychol Assoc. 2003;22(5):487–97.
8. Tedeschi RG, Calhoun LG. Trauma & transformation: growing in the aftermath of suffering. Thousand Oaks: Sage; 1995. p. 163.
9. Tedeschi RG, Calhoun LG. The posttraumatic growth inventory: measuring the positive legacy of trauma. J Trauma Stress. 1996;9(3):455–71.
10. Tedeschi RG, Shakespeare-Finch J, Taku K, Calhoun LG. Posttraumatic growth: theory, research and applications. New York London: Routledge, Taylor and Francis Group; 2018. p. 256.
11. Lelorain S, Bonnaud-Antignac A, Florin A. Long term posttraumatic growth after breast Cancer: prevalence, predictors and relationships with psychological health. J Clin Psychol Med Settings. 2010;17(1):14–22.
12. Cadell S, Suarez E, Hemsworth D. Reliability and validity of a French version of the posttraumatic growth inventory. Open J Med Psychol. 2015;04(02):53–65.
13. Magne H, Jaafari N, Voyer M. La croissance post-traumatique : un concept méconnu de la psychiatrie française. L'Encéphale. 2021;47(2):143–50.
14. Bellizzi KM, Smith AW, Reeve BB, Alfano CM, Bernstein L, Meeske K, et al. Posttraumatic growth and health-related quality of life in a racially diverse cohort of breast Cancer survivors. J Health Psychol. 2010;15(4):615–26.
15. Da Silva SIM, Da Cruz Moreira HT, De Aguiar Pinto SM, Portocarrero Canavarro MCCS. Cancro da mama e desenvolvimento pessoal e relacional: Estudo das características psicométricas do Inventário de Desenvolvimento Pós-Traumático (posttraumatic growth inventory) numa amostra de mulheres da população Portuguesa. [breast cancer and personal and relational growth: psychometric characteristics of the Portuguese version of the posttraumatic growth inventory in a sample of Portuguese women.]. Rev Iberoam Diagnóstico Eval Psicol. 2009;28(2):105–33.
16. Ho SMY, Chan CLW, Ho RTH. Posttraumatic growth in chinese cancer survivors. Psychooncology. 2004;13(6):377–89.
17. Powell S, Rosner R, Butollo W, Tedeschi RG, Calhoun LG. Posttraumatic growth after war: a study with former refugees and displaced people in Sarajevo. J Clin Psychol. 2003;59(1):71–83.
18. Rodríguez-Rey R, Alonso-Tapia J, Kassam-Adams N. The factor structure of the posttraumatic growth inventory in parents of critically ill children. Psicothema. 2016;(28.4):495–503.
19. Taku K, Calhoun LG, Tedeschi RG, Gil-Rivas V, Kilmer RP, Cann A. Examining posttraumatic growth among Japanese university students. Anxiety Stress Coping. 2007;20(4):353–67.
20. Weiss T, Berger R. Reliability and validity of a Spanish version of the posttraumatic growth inventory. Res Soc Work Pract. 2006;16(2):191–9.
21. Danhauer SC, Case LD, Tedeschi R, Russell G, Vishnevsky T, Triplett K, et al. Predictors of posttraumatic growth in women with breast cancer. Psychooncology. 2013;22(12). https://doi.org/10.1002/pon.3298.
22. Danhauer SC, Russell GB, Tedeschi RG, Jesse MT, Vishnevsky T, Daley K, et al. A longitudinal investigation of posttraumatic growth in adult patients undergoing treatment for acute leukemia. J Clin Psychol Med Settings. 2013;20(1):13–24.

23. Danhauer SC, Russell G, Case LD, Sohl SJ, Tedeschi RG, Addington EL, et al. Trajectories of posttraumatic growth and associated characteristics in women with breast Cancer. Ann Behav Med Publ Soc Behav Med. 2015;49(5):650–9.
24. Brédart A, Untas A, Copel L, Leufroy M, Mino JC, Boiron C, et al. Breast Cancer survivors' supportive care needs, posttraumatic growth and satisfaction with doctors' interpersonal skills in relation to physical activity 8 months after the end of treatment: a prospective exploratory study. Oncology. 2016;90(3):151–9.
25. Porro B, Michel A, Zinzindohoué C, Bertrand P, Monrigal E, Trentini F, et al. Prise en charge psychologique des femmes ayant un cancer du sein. Quelles différences interindividuelles sur le développement post-traumatique durant la première année suivant le diagnostic ? Psycho-Oncol. 2019;13(3–4):168–72.
26. Blanchin M, Dauchy S, Cano A, Brédart A, Aaronson NK, Hardouin JB. Validation of the French translation-adaptation of the impact of cancer questionnaire version 2 (IOCv2) in a breast cancer survivor population. Health Qual Life Outcomes. 2015;13:110.
27. Bourdon M, Blanchin M, Campone M, Quéreux G, Dravet F, Sébille V, et al. A comparison of posttraumatic growth changes in breast cancer and melanoma. Health Psychol. 2019;38(10):878–87.
28. Bouhnik AD, Bendiane MK, Cortaredona S, Sagaon Teyssier L, Rey D, Berenger C, et al. The labour market, psychosocial outcomes and health conditions in cancer survivors: protocol for a nationwide longitudinal survey 2 and 5 years after cancer diagnosis (the VICAN survey). BMJ OpenBMJ Open. 2015;5(3):e005971
29. Colombié M, Jézéquel P, Rubeaux M, Frenel JS, Bigot F, Seegers V, et al. The EPICURE study: a pilot prospective cohort study of heterogeneous and massive data integration in metastatic breast cancer patients. BMC Cancer. 2021;21:333.
30. Zhou K, Blanc-Lapierre A, Seegers V, Boisdron-Celle M, Bigot F, Bourdon M, et al. Anosmia but not Ageusia as a COVID-19-related symptom among Cancer patients—first results from the PAPESCO-19 cohort study. Cancers. 2021;13(14):3389.
31. Sébille V, Hardouin JB, Giral M, Bonnaud-Antignac A, Tessier P, Papuchon E, et al. Prospective, multicenter, controlled study of quality of life, psychological adjustment process and medical outcomes of patients receiving a preemptive kidney transplant compared to a similar population of recipients after a dialysis period of less than three years – The PreKit-QoL study protocol. BMC Nephrol. 2016;17(1). [cited 2019 Jul 29] Available from: http://www.biomedcentral.com/1471-2369/17/11
32. Corman M, Rubio MT, Cabrespine A, Brindel I, Bay JO, De La Tour RP, et al. Retrospective and prospective measures of post-traumatic growth reflect different processes: longitudinal evidence of greater decline than growth following a hematopoietic stem-cell transplantation. BMC Psychiatry. 2021;21(1):27.
33. Paunescu AC, Copie CB, Malak S, Gouill SL, Ribrag V, Bouabdallah K, et al. Quality of life of survivors 1 year after the diagnosis of diffuse large B-cell lymphoma: a LYSA study. Ann Hematol. 2021.
34. Teixeira RJ, Pereira MG. Growth and the cancer caregiving experience: psychometric properties of the Portuguese posttraumatic growth inventory. Fam Syst Health. 2013;31(4):382–95.
35. Bourdon M, Blanchin M, Tessier P, Campone M, Quéreux G, Dravet F, et al. Changes in quality of life after a diagnosis of cancer: a 2-year study comparing breast cancer and melanoma patients. Qual Life Res. 2016;25(8):1969–79.
36. Carver CS. You want to measure coping but your protocol' too long: consider the brief cope. Int J Behav Med. 1997;4(1):92–100.
37. Muller L, Spitz E. Multidimensional assessment of coping: validation of the brief COPE among French population. L'Encephale. 2003;29(6):507–18.
38. Carver CS, Scheier MF, Weintraub JK. Assessing coping strategies: a theoretically based approach. J Pers Soc Psychol. 1989;56(2):267–83.
39. Little TD, editor. The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2 [Internet]. 1st ed. Oxford University Press; 2013 [cited 2022 Jul 10]. Available from: http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199934898.001.0001/oxfordhb-9780199934898
40. Taku K, Cann A, Calhoun LG, Tedeschi RG. The factor structure of the posttraumatic growth inventory: a comparison of five models using confirmatory factor analysis. J Trauma Stress. 2008;21(2):158–64.
41. Brunet J, McDonough MH, Hadd V, Crocker PRE, Sabiston CM. The post-traumatic growth inventory: an examination of the factor structure and invariance among breast cancer survivors. Psychooncology. 2010;19(8):830–8.
42. Jaarsma TA, Pool G, Sanderman R, Ranchor AV. Psychometric properties of the Dutch version of the posttraumatic growth inventory among cancer patients: Dutch version of the posttraumatic growth inventory. Psychooncology. 2006;15(10):911–20.
43. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297–334.
44. Mesbah M. Statistical Quality of Life. In: Methods and Applications of Statistics in the Life and Health Sciences. Balakrishnan, N., Read, C.B., Vidakovic, B., Kotz, S., Johnson, N.L. Wiley; 2010.
45. Sijtsma K, Molenaar IW. Introduction to nonparametric item response theory. Thousand Oaks, Calif.: SAGE; 2002. 167 p. (Meaurement methods for the social sciences).
46. Ark LA van der. Mokken Scale Analysis in *R*. J Stat Softw 2007 [cited 2022 Jul 10];20(11). Available from: http://www.jstatsoft.org/v20/i11/
47. Vigneau E, Qannari EM. Clustering of variables around latent components. Commun Stat - Simul Comput. 2003;32(4):1131–50.
48. Tedeschi RG, Calhoun LG. Posttraumatic growth: conceptual foundations and empirical evidence. Psychol Inq. 2004;15(1):1–18.
49. Lehto US, Ojanen M, Kellokumpu-Lehtinen P. Predictors of quality of life in newly diagnosed melanoma and breast cancer patients. Ann Oncol. 2005;16(5):805–16.
50. Henson C, Truchot D, Canevello A. What promotes post traumatic growth? A systematic review. Eur J Trauma Dissociation. 2021;5(4):100195.
51. Hardouin JB. CLV: Stata module to implement a clustering of variables around latent components [Internet]. Statistical Software Components. Boston College Department of Economics; 2019 [cited 2021 Oct 18]. Available from: https://ideas.repec.org/c/boc/bocode/s453101.html
52. Bohlmeijer ET, Lamers SMA, Fledderus M. Flourishing in people with depressive symptomatology increases with acceptance and commitment therapy. Post-hoc analyses of a randomized controlled trial. Behav Res Ther. 2015;65:101–6.
53. Fledderus M, Bohlmeijer ET, Smit F, Westerhof GJ. Mental health promotion as a new goal in public mental health care: a randomized controlled trial of an intervention enhancing psychological flexibility. Am J Public Health. 2010;100(12):2372.
54. Wersebe H, Lieb R, Meyer AH, Hofer P, Gloster AT. The link between stress, well-being, and psychological flexibility during an acceptance and commitment therapy self-help intervention. Int J Clin Health Psychol. 2018;18(1):60–8.
55. Cormio C, Romito F, Giotta F, Mattioli V. Post-traumatic growth in the Italian experience of long-term disease-free Cancer survivors. Stress Health. 2015;31(3):189–96.
56. Morris BA, Shakespeare-Finch J, Scott JL. Coping processes and dimensions of posttraumatic growth. Australas J Disaster Trauma Stud. 2007;2007(1):No Pagination Specified-No Pagination Specified.
57. Schmidt SD, Blank TO, Bellizzi KM, Park CL. The relationship of coping strategies, social support, and attachment style with posttraumatic growth in cancer survivors. J Health Psychol. 2012 Oct;17(7):1033–40.
58. Abu-Raiya H, Pargament KI, Mahoney A. Examining coping methods with stressful interpersonal events experienced by Muslims living in the United States following the 9/11 attacks. Psychol Relig Spiritual. 2011;3(1):1–14.
59. Gerber MM, Boals A, Schuettler D. The unique contributions of positive and negative religious coping to posttraumatic growth and PTSD. Psychol Relig Spiritual. 2011;3(4):298–307.
60. Pargament KI, Magyar GM, Benore E, Mahoney A. Sacrilege: a study of sacred loss and desecration and their implications for health and well-being in a community sample. J Sci Study Relig. 2005;44(1):59–78.
61. Bourdon M, Bonnaud-Antignac A, Roussiau N, Quéreux G. Spiritualité et changement de valeurs chez des patients atteints d'un mélanome: une étude qualitative exploratoire. Psycho-Oncol. 2011;5(1):34–9.
62. Cordova M, Giese-Davis J, Golant M, Kronenwetter C, Chang V, Spiegel D. Breast Cancer as trauma: posttraumatic stress and posttraumatic growth. J Clin Psychol Med Settings. 2007;14:308–19.

63. Lechner SC, Zakowski SG, Antoni MH, Greenhawt M, Block K, Block P. Do sociodemographic and disease-related variables influence benefit-finding in cancer patients? Psychooncology. 2003;12(5):491–9.
64. Tanyi Z, Szluha K, Nemes L, Kovács S, Bugán A. Positive consequences of Cancer: exploring relationships among posttraumatic growth, adult attachment, and quality of life. Tumori J. 2015;101(2):223–31.
65. Cole BS, Hopkins CM, Tisak J, Steel JL, Carr BI. Assessing spiritual growth and spiritual decline following a diagnosis of cancer: reliability and validity of the spiritual transformation scale. Psychooncology. 2008;17(2):112–21.
66. Davis LZ, Cuneo M, Thaker PH, Goodheart MJ, Bender D, Lutgendorf SK. Changes in spiritual well-being and psychological outcomes in ovarian cancer survivors. Psychooncology. 2018;27(2):477–83.
67. Cungi C. L'alliance thérapeutique. Retz; 2006. 286 p.
68. Gori A, Topino E, Sette A, Cramer H. Pathways to post-traumatic growth in cancer patients: moderated mediation and single mediation analyses with resilience, personality, and coping strategies. J Affect Disord. 2021;279:692–700.

Publisher's Note

Supplementary Materials

Manuscript title:

Posttraumatic Growth Inventory: challenges with its validation among French cancer patients

Appendix A: Inventory of the wording changes between the French translation realized by Lelorain et al. (PTGI French version 1) and the version that derived from this translation (version 2)

We noticed two French versions of the PTGI equally used in France (see Table 1 of our manuscript):

- **Version 1 (V1):** The original translation realized by Lelorain et al.

- **Version 2 (V2):** A revised version with slight adaptations in wording compared to the original translation.

The revised version V2 seemed to have appeared in 2010 following the work of Lelorain et al. It first appeared in a French study protocol. Since then, this revised version has been disseminated in France to various research teams without any mention of the changes made to the wording of some items.

We noted changes in the wording of three items when comparing different study protocols involving the PTGI (the remaining 18 items being identical). These wording changes are addressed in Table S1.

276

Table S1: Inventory of the wording changes observed between the original translation of Lelorain et al. (V1) and the revised version (V2)

Item with wording changes		Authors' comments
13.	V1 : J'apprécie **davantage** chaque jour de ma vie V2 : J'apprécie **plus amplement** chaque jour de ma vie	This is a minor change, as the meaning of the item is maintained.
14.	V1 : De nouvelles opportunités sont apparues, qui ne seraient pas apparues autrement V2 : De nouvelles opportunités sont apparues, ~~qui ne seraient pas apparues autrement~~	The end of the item (indicating that the new opportunities would not have appeared without the disease) has been deleted in the version 2. Item from version 1 is the closest to the English version.
20.	V1 : J'ai vraiment compris à quel point les gens pouvaient être formidables. V2 : Je vois plus le bon côté des gens	Item from version 1 is a literal translation of the English version. It is semantically correct, but it may not be adapted to the French culture. The item has been entirely rephrased in version 2, probably to overcome this issue. If the notion of goodness in people can be found in both wording, it is much more prominent in the original version.

Appendix B: French version of the posttraumatic growth inventory used in the ELCCA study and evaluated in our manuscript. It corresponds to the revised version (i.e., version 2) derived from the translation of Lelorain et al. The wording of the items in the English version is indicated in square brackets.

Inventaire du Développement Post-Traumatique

DU FAIT DE MON CANCER :	PAS DU TOUT 0	TRÈS PEU 1	PEU 2	UN PEU 3	BEAUCOUP 4	TOTALEMENT 5
1. J'ai changé de priorités dans la vie [I changed my priorities about what is important in life]						
2. J'apprécie plus ma vie à sa vraie valeur [I have a greater appreciation for the value of my own life]						
3. Je me suis intéressé(e) à de nouvelles choses [I developed new interests]						
4. J'ai acquis plus confiance en moi [I have a greater feeling of self-reliance]						
5. J'ai développé une certaine spiritualité [I have a better understanding of spiritual matters]						
6. Je vois mieux que je peux compter sur les autres en cas de problème [I more clearly see that I can count on people in times of trouble]						
7. J'ai donné une nouvelle direction à ma vie [I established a new path for my life]						
8. Je me sens plus proche des autres [I have a greater sense of closeness with others]						
9. Je suis plus enclin(e) à exprimer mes émotions [I have a greater willingness to express my emotions]						
10. Je suis davantage capable de gérer des situations difficiles [I know better that I can handle difficulties]						
11. Je fais de ma vie quelque chose de meilleur [I'm able to do better things with my life]						
12. J'accepte mieux la façon dont les choses se passent [I am better able to accept the way things work out]						
13. J'apprécie plus amplement chaque jour de ma vie [I can better appreciate each day]						
14. De nouvelles opportunités sont apparues [New opportunities are available which wouldn't have been otherwise]						
15. J'ai plus de compassion pour les autres [I have greater compassion for others]						
16. J'investis plus mes relations aux autres [I put more effort into my relationships]						
17. J'essaie davantage de changer les choses qui ont besoin d'être changées [I'm more likely to try to change things which need changing]						
18. J'ai une foi religieuse plus grande [I have a stronger religious faith]						
19. J'ai découvert que je suis plus fort(e) que ce que je pensais [I discovered that I'm stronger than I thought I was]						
20. Je vois plus le bon côté des gens [I learned a great deal about how wonderful people are]						
21. J'accepte mieux le fait d'avoir besoin des autres [I better accept needing others]						

Appendix C: Standardized factor loadings from the confirmatory factor analysis based on the newly evidenced four-factor structure

Table S2: Standardized factor loadings of the 21 items of the posttraumatic growth inventory obtained after the oblique confirmatory factor analysis based on four-factor structure evidenced by the Clustering around Latent Variable analysis

	Standardized factor loadings for each factor			
PTGI items	RO	NLD	PC	SC
Relating to others (RO)				
6. I more clearly see that I can count on people in times of trouble *Je vois mieux que je peux compter sur les autres en cas de problème*	0.55	-	-	-
8. I have a greater sense of closeness with others *Je me sens plus proche des autres*	0.82	-	-	-
9. I have a greater willingness to express my emotions *Je suis plus enclin(e) à exprimer mes émotions*	0.75	-	-	-
15. I have greater compassion for others *J'ai plus de compassion pour les autres*	0.63	-	-	-
16. I put more effort into my relationships *J'investis plus mes relations aux autres*	0.80	-	-	-
20. I learned a great deal about how wonderful people are *Je vois plus le bon côté des gens*	0.75	-	-	-
21. I better accept needing others *J'accepte mieux le fait d'avoir besoin des autres*	0.69	-	-	-
New life orientation (NLD)				
3. I developed new interests *Je me suis intéressé(e) à de nouvelles choses*	-	0.78	-	-
7. I established a new path for my life *J'ai donné une nouvelle direction à ma vie*	-	0.80	-	-
11. I'm able to do better things with my life *Je fais de ma vie quelque chose de meilleur*	-	0.82	-	-
14. New opportunities are available which wouldn't have been otherwise *De nouvelles opportunités sont apparues*	-	0.65	-	-
17. I'm more likely to try to change things which need changing *J'essaie davantage de changer les choses qui ont besoin d'être changées*	-	0.71	-	-
1. I changed my priorities about what is important in life *J'ai changé de priorités dans la vie*	-	0.63	-	-
Personal capacities (PC)				
4. I have a greater feeling of self-reliance *J'ai acquis plus confiance en moi*	-	-	0.73	-
10. I know better that I can handle difficulties *Je suis davantage capable de gérer des situations difficiles*	-	-	0.75	-
12. I am better able to accept the way things work out *J'accepte mieux la façon dont les choses se passent*	-	-	0.78	-
19. I discovered that I'm stronger than I thought I was *J'ai découvert que je suis plus fort(e) que ce que je pensais*	-	-	0.65	-
2. I have a greater appreciation for the value of my own life *J'apprécie plus ma vie à sa vraie valeur*	-	-	0.75	-
13. I can better appreciate each day *J'apprécie plus amplement chaque jour de ma vie*	-	-	0.80	-
Spiritual change (SC)				
5. I have a better understanding of spiritual matters *J'ai développé une certaine spiritualité*	-	-	-	0.93
18. I have a stronger religious faith *J'ai une foi religieuse plus grande*	-	-	-	0.83

Appendix D: The final French version of the posttraumatic growth inventory that we propose, with five response categories per item and four domains (resulting from the analyses of the manuscript)

Inventaire du Développement Post-Traumatique

DU FAIT DE MON CANCER :	PAS DU TOUT 0	TRÈS PEU 1	UN PEU 2	BEAUCOUP 3	TOTALEMENT 4
1. J'ai changé de priorités dans la vie (NLD)					
2. J'apprécie plus ma vie à sa vraie valeur (PC)					
3. Je me suis intéressé(e) à de nouvelles choses (NLD)					
4. J'ai acquis plus confiance en moi (PC)					
5. J'ai développé une certaine spiritualité (SC)					
6. Je vois mieux que je peux compter sur les autres en cas de problème (RO)					
7. J'ai donné une nouvelle direction à ma vie (NLD)					
8. Je me sens plus proche des autres (RO)					
9. Je suis plus enclin(e) à exprimer mes émotions (RO)					
10. Je suis davantage capable de gérer des situations difficiles (PC)					
11. Je fais de ma vie quelque chose de meilleur (NLD)					
12. J'accepte mieux la façon dont les choses se passent (PC)					
13. J'apprécie plus amplement chaque jour de ma vie (PC)					
14. De nouvelles opportunités sont apparues (NLD)					
15. J'ai plus de compassion pour les autres (RO)					
16. J'investis plus mes relations aux autres (RO)					
17. J'essaie davantage de changer les choses qui ont besoin d'être changées (NLD)					
18. J'ai une foi religieuse plus grande (SC)					
19. J'ai découvert que je suis plus fort(e) que ce que je pensais (PC)					
20. Je vois plus le bon côté des gens (RO)					
21. J'accepte mieux le fait d'avoir besoin des autres (RO)					

Notes. Items are highlighted according to their domain (domains are also shown for each item in brackets): NLD: New life direction / Nouvelle direction pour la vie (6 items), PC: Personal capacities / Capacités personelles (6 items), SC: Spiritual Chance / Changement spirituel (2 items), RO: Relating to others / Relation aux autres (7 items)

6.5 Bilan

Les travaux présentés dans ce chapitre visaient à étudier les propriétés psychométriques d'une des traductions françaises de l'inventaire du développement post-traumatique (PTGI). Lorsque nous avons initié ces travaux, nous savions que deux traductions existaient :

- Une traduction française (réalisée par Sophie Lelorain *et al.* [182, 183])
- Une traduction en français canadien (réalisée par Cadell *et al.* [191]).

Nous nous sommes ensuite rendus compte que la version du questionnaire dont nous disposions ne correspondait pas totalement à la traduction proposée par Lelorain *et al.* [182] qui n'avait, à l'époque, pas été publiée. Cette version dérivée semble avoir fait son apparition avec l'étude ELCCA (une étude longitudinale s'intéressant aux changements psychologiques et socio-économiques à la suite d'un diagnostic de cancer du sein ou de mélanome) [193], dont la première inclusion a eu lieu en novembre 2010.

Afin de dresser un état des lieux de l'utilisation du PTGI en France, nous avons cherché à recenser les études françaises s'intéressant au développement post-traumatique et impliquant une mesure grâce au PTGI. Pour ce faire, nous avons recoupé différentes sources d'informations : manuscrits de thèse de doctorat, *case report form* et protocoles d'études, articles scientifiques et livres de psychologie. Cet état des lieux n'est malheureusement probablement pas exhaustif. De plus, il a été complexifié par le fait que certains auteurs citaient la traduction de Cadell *et al.* [191], alors que leur étude impliquait en fait une autre version (celle de Lelorain *et al.* [version 1] [182], ou la version qui en a dérivé [version 2]). Cet état des lieux nous a cependant permis de remarquer que les versions les plus utilisées en France étaient les versions 1 et 2, même si la version de Cadell *et al.* [191] (version 3) apparait également dans quelques travaux.

Face à la multiplicité des versions françaises (et les différences de formulation d'item observées), nous nous sommes tournés vers la formulation initiale des items de Tedeschi et Calhoun [24] et nous avons remarqué que plusieurs items n'avaient plus tout à fait le même sens après avoir été traduits. De plus, le format des modalités de réponse des versions 1 et 2 (0 = "pas du

tout", 1 = "très peu", 2 = "peu", 3 = "un peu", 4 = "beaucoup", 5 = "totalement") est différent du format de modalité de réponse initial (allant de 0 = "*I did not experience this change*" à 5 = "*I experienced this change to a very great degree*"). La problématique est qu'en français, les modalités "Très peu", "Peu", et "Un peu" sont très proches, et il nous a semblé qu'elles pourraient dérouter les patients.

L'étude des propriétés psychométriques de la version française dérivée de la traduction de Lelorain *et al.* (version 2) a été réalisée grâce aux données de la cohorte ELCCA, en se plaçant à deux ans post diagnostic. Une étude confirmatoire (cherchant à valider la structure à cinq dimensions proposée par Tedeschi et Calhoun [24]) a mis en évidence une dimension problématique : la dimension "Appréciation de la vie". Les items de cette dimension étaient plus corrélés avec d'autres dimensions que leur dimension hypothétique. De plus, un problème de dimensionnalité a été mis en évidence par la courbe alpha de Cronbach. L'item qui semblait être particulièrement problématique dans cette dimension était l'item n°1 "J'ai changé de priorité dans la vie". Une analyse exploratoire de type *clustering around latent variable* (CLV) a permis de mettre en évidence une structure à quatre dimensions, assez proche de la structure initiale, et qui semblait présenter de meilleures propriétés. Les deux dimensions qui différaient par rapport à la version initiale ont été renommées pour mieux refléter leur contenu. Il s'agit des dimensions :

- "Nouvelle direction pour la vie" (qui regroupe les items de la dimension originelle "Nouvelles opportunités" et l'item n°1 s'intéressant aux nouvelles priorités de vie) ;
- "Capacités personnelles" (qui regroupe les items de la dimension originelle "Force personnelle" et les items n°2 et n°13 d'appréciation de la valeur de la vie et du quotidien).

Avant de réaliser l'analyse CLV, nous avons décidé de regrouper les modalités de réponse "très peu" et "peu". Ce choix a été motivé par l'estimation de cinq modèles de crédit partiel (un pour chacune des cinq dimensions) grâce auxquels nous avons pu observer un chevauchement entre les courbes caractéristiques des modalités de réponse "très peu" et "peu" (ces résultats ne sont pas évoqués dans l'article).

Lors de nos analyses, nous avons pu remarquer que les scores à la dimension "Changement spirituel" étaient très bas. En effet, le score médian était de 1,0 (écart interquartile : 3,0) chez les patientes atteintes d'un cancer du sein et de 0,0 (écart interquartile : 1,5) chez les patient(e)s atteint(e)s d'un mélanome. Si cela s'explique en partie par le faible nombre d'items dans la dimension, Tedeschi *et al.* ont également noté qu'un effet plancher pouvait être observé pour cette dimension dans les pays laïques [186] (comme c'est le cas de la France) . Pour avoir une dimension adaptée à cette différence culturelle par rapport aux États-Unis, une extension du PTGI a été proposée : le PTGI-X [194]. Cette extension contient les 21 items originaux auxquels ont été ajoutés quatre nouveaux items s'intéressant aux changements existentiels (mieux comprendre le sens de la vie, être capable d'affronter les questionnements concernant la vie et la mort, se sentir plus en harmonie avec le monde, etc.). Ces items ont été regroupés avec les items liés aux changements spirituels au sein d'une seule et même dimension dénommée "Changements existentiels et spirituels". Il pourrait être intéressant d'intégrer ces items dans le questionnaire français.

Trois autres études sur le PTGI ont été menées contemporainement à nos travaux. Elles ne sont pas mentionnées dans l'article présenté dans la section précédente, car elles n'étaient pas encore publiées où nous l'avons soumis. Il s'agit des études de Porro *et al.* [195] de Evans *et al.* [196] et de Henson *et al.* [197] :

- Porro *et al.* [195] ont cherché à valider une version courte du PTGI chez des patientes atteintes d'un cancer du sein. Cette version courte s'appuie sur la traduction de Lelorain *et al.* [182]. Ces auteurs ont mis en évidence deux versions courtes à cinq dimensions présentant des propriétés psychométriques satisfaisantes (parmi les trois versions courtes investiguées). Ces versions courtes sont issues des travaux de Cann *et al.* [198], Prati et Pietrantoni [199] et Kaur *et al.* [200].

- Evans *et al.* [196] ont publié des résultats issus de l'étude VICAN, une étude nationale qui s'intéresse au devenir des patients après le diagnostic de leur cancer. Ces auteurs ont plus particulièrement cherché à identifier les facteurs influençant le développement post-

traumatique. Cette étude s'appuie sur la version n°2 (c'est-à-dire celle que l'on a étudiée).

- Henson *et al.* ont réalisé une nouvelle traduction du PTGI en français [197], dont ils ont étudié les propriétés psychométriques à partir d'un échantillon de pompiers. Ils justifient cette nouvelle traduction en indiquant que :

  (i) la traduction de Cadell *et al.* [191] pourrait ne pas être adaptée linguistiquement et culturellement à une utilisation en France ;

  (ii) la traduction de Lelorain *et al.* [182] pourrait être peu fiable (car traduite par les auteures elles-mêmes et présentant un mauvais ajustement à la structure initiale de Tedeschi et Calhoun [24]) ;

  (iii) les chercheurs français ne peuvent actuellement pas étudier quantitativement le développement post-traumatique.

  Dans les faits, ce troisième point est discutable, puisque nombre d'études françaises ont déjà présenté des résultats sur les scores issus d'une des versions du PTGI. Henson *et al.* ont donc traduit le questionnaire (eux-mêmes) et ont fait appel à un traducteur certifié pour réaliser une rétro-traduction. Aucune différence substantielle n'a été mise en évidence entre la version initiale et la rétro-traduction. Avec cette nouvelle version du questionnaire, les cinq dimensions originelles présentaient toutes une bonne consistance interne (alpha de Cronchach supérieur à 0,7), excepté la dimension "Changement spirituel" (α = 0,6). Leur analyse factorielle confirmatoire, basée sur la structure à cinq facteurs proposée par Tedeschi et Calhoun [184], présentait également un ajustement acceptable. Cependant, ces auteurs indiquent qu'ils ont ajouté des corrélations entre les résidus du modèle, ce que n'avait pas fait Lelorain *et al.* [182]). Henson *et al.* concluaient néanmoins sur le fait qu'ils pensaient que le PTGI devrait être repensé et amélioré [197]. En étudiant cette nouvelle traduction, on peut remarquer que les items n°5, 11, 12 et 13 ont une formulation qui semble plus proche de la version de Tedeschi et Calhoun que la traduction que nous avons étudiée.

Épilogue

Notre étude, basée sur les critères de COSMIN [201, 202], a permis de fournir des données quant aux propriétés psychométriques d'une des versions du PTGI (fiabilité, validité de structure, validité concourante, et validité de type *known-group*). À noter qu'aucune information sur la validité de face des versions françaises du PTGI n'est actuellement disponible (pour aucunes des versions). Elle aurait dû être éprouvée lors du développement des traductions françaises (avant qu'elles ne soient diffusées à "grande échelle", comme c'est le cas aujourd'hui). Cependant, ni Lelorain *et al.* [182] ni Cadell *et al.* [191] n'en font mention. Il en va de même pour la nouvelle traduction de Henson *et al.* [197].

Il serait bénéfique, pour la recherche et la pratique clinique, qu'une réflexion commune sur ce questionnaire s'organise afin de délivrer des recommandations claires sur son utilisation en France.

Discussion générale

Discussion générale

Cette section est polarisée autour des trois chapitres relatifs aux travaux de thèse. Pour chaque chapitre, un bilan et les perspectives directes sont dressées.

Individualisation de la détection du *response shift* à l'aide des erreurs de Guttman

Les travaux de cette thèse visaient initialement à individualiser la détection du *response shift* lors de l'analyse des données rapportées par les patients entre deux temps de mesure. La méthode de détection que nous avons étudiée est basée sur une proposition de Blanchin *et al.* [148]. Elle consistait à étudier l'évolution du nombre d'erreurs de Guttman pour classer les individus en deux groupes :

Groupe 1 : Individus présentant une augmentation importante du nombre d'erreurs de Guttman (supposés susceptibles d'expérimenter du *response shift*) ;

Groupe 2 : Individus présentant un nombre d'erreurs de Guttman globalement stable (supposés peu susceptibles d'être affectés par du *response shift*).

Cette méthode a été formalisée dans le chapitre 4 grâce à l'introduction d'indicateurs traduisant l'évolution du nombre d'erreurs de Guttman entre les deux temps de mesure (les indicateurs I et I_{norm}). Afin de déterminer si l'hypothèse de cette méthode était valide, nous avons mené des études de simulation de type "preuve de concept". L'objectif était de confirmer les hypothèses suivantes :

1. Les individus fictifs pour lesquels du *response shift* est simulé présentent effectivement une augmentation du nombre d'erreurs de Guttman entre les deux temps de mesure (correspondant à des valeurs d'indicateurs supérieures à 0) ;
2. Les individus fictifs pour lesquels aucun *response shift* n'est simulé présentent un nombre d'erreurs de Guttman stable dans le temps (correspondant à des valeurs d'indicateurs proches de 0).

Dans nos simulations, nous sommes également intéressés au chevauchement des distributions des indicateurs chez les individus pour qui on avait simulé du *response shift* et chez les individus pour qui on n'en avait pas simulé. L'objectif était de quantifier la capacité discriminante de ces indicateurs (c'est-à-dire leur capacité à distinguer les individus affectés par le *response shift* des individus non affectés).

Au cours de ces travaux, nous nous sommes cantonnés au cas d'un questionnaire unidimensionnel et nous n'avons considéré qu'une seule forme de *response shift* : la recalibration. Nous n'avons pas considéré la reconceptualisation, car elle aurait nécessité une modalisation multidimensionnelle. Quant à la repriorisation, si elle est facilement conceptualisée au niveau des dimensions (exemple du fonctionnement social devenant plus important après un accident de voiture qui entraine des séquelles physiques importantes), il nous semble moins évident qu'un item devienne plus important que d'autres au cours du temps dans un contexte unidimensionnel [149].

Après avoir réalisé plusieurs études de simulation, nous avons constaté que la recalibration s'accompagnait effectivement d'une légère augmentation du nombre d'erreurs de Guttman. Cet écart n'était néanmoins pas discriminant, ne permettant pas de s'appuyer sur les erreurs de Guttman pour une détection individuelle du *response shift*. Le même type de résultat a été obtenu avec des indicateurs basés sur des indices de *fit* paramétriques comme l'INFIT et l'OUTFIT.

Ces résultats rappellent la difficulté de faire de la statistique à un niveau individuel. D'autres méthodes statistiques pourraient permettre d'étudier la variabilité interindividuelle du *response shift* et tendre vers une granularité plus individuelle. On peut par exemple citer les *Latent Variable Mixture Models* (modèles à variables latente de mélange). Lors de premiers tests réalisés au cours de cette thèse, l'estimation d'un PCM longitudinal de mélange s'est avérée extrêmement longue. En effet, l'estimation d'un seul modèle nécessitait plusieurs heures de calculs. De plus, ces modèles sont connus pour converger sur des *extremums* locaux. Ces constats compliquent la mise en place d'une étude de simulation et l'implémentation pratique de ce type de modèles. Une autre méthode dont on ne connait pas les performances consiste à étudier les trajectoires des résidus d'un modèle

linéaire mixte [151]. Cette méthode, présentée dans le chapitre 3, pourrait également être évaluée par simulations. L'objectif serait alors de déterminer si :

- Le *response shift* se manifeste effectivement par une fluctuation importante des résidus centrés chez les individus qui l'expérimentent ;
- Les individus n'expérimentant pas de *response shift* (c.-à-d., ceux pour qui on n'en a pas simulé) ont bien des résidus centrés qui fluctuent autour de 0.

L'intérêt de ces deux méthodes réside dans le fait qu'elles ne nécessitent pas l'introduction de covariables supposées expliquer la survenue de *response shift*.

Comme on l'a vu dans l'état des connaissances (chapitre 3), une méthode permettant la détection du *response shift* à un niveau individuel consiste à avoir recours à des entretiens semi-dirigés (méthode qualitative). Des chercheurs en psychologie à SPHERE ont mené de premiers essais en ce sens dans le cadre du projet HAP-2 [1]. Cependant, lors de l'élaboration des questions de l'entretien, la notion de recalibration a semblé très difficile à appréhender, et n'a finalement pas été retenue. Des méthodes mixtes (regroupant à la fois approches qualitatives et quantitatives) pourrait être envisagées pour dresser un portrait plus complet des changements expérimentés par les patients au cours du temps. Ce type de méthode pourrait permettre de tendre vers une détection plus individualisée, grâce à l'analyse de l'évolution des réponses des individus éclairée par les informations d'entretiens qualitatifs.

Quelles que soient les orientations que prend la recherche pour la détection du *response shift* au niveau individuel, il sera capital de considérer les potentielles explications alternatives. En effet, pour toutes les méthodes de détection, des phénomènes autres que le *response shift* peuvent conduire à la conclusion que du *response shift* est survenu, alors que ce n'est en fait pas le cas. Sébille et al. ont récemment dressé un panorama des explications alternatives propres à chaque méthode [23].

1. *HAP-2 est un projet qui vise à améliorer la prévention et le traitement des pneumonies nosocomiales.*

Extension de la partie 1 de l'algorithme ROSALI

Une autre piste pour prendre en compte l'hétérogénéité du *response shift* consiste à introduire des covariables dans le processus de détection. L'algorithme ROSALI, développé au sein de l'unité SPHERE, a été récemment étendu en ce sens avec les travaux de Hammas *et al.* [150].

Lorsque mes travaux de thèse ont été initiés, l'algorithme ROSALI était déjà basé sur la théorie de la mesure de Rasch (estimation de PCM : ROSALI-RMT) et permettait de détecter de la recalibration entre deux temps de mesure en prenant en compte une covariable binaire (la recalibration étant opérationnalisée comme un changement longitudinal dans les valeurs des paramètres de seuil des items). Dans cette version à une covariable, l'algorithme cherche tout d'abord à identifier s'il existe des différences dans les paramètres de seuils des items au premier temps de mesure entre les deux groupes d'individus définis par la covariable (partie 1). Une fois ces potentielles différences investiguées, l'algorithme recherche ensuite de la présence de recalibration entre les deux temps de mesure (partie 2). Cette recalibration peut être commune aux deux groupes ou différentielle (c'est-à-dire survenir dans un seul des deux groupes, ou survenir de façon différente dans les deux groupes).

Une partie de mes travaux de thèse visaient à continuer d'étendre cet algorithme, pour non plus prendre en compte une seule, mais deux covariables binaires (qui pourraient correspondre à des caractéristiques différentes des répondants, comme le sexe et une caractéristique clinique). Ces travaux de thèse ont porté sur l'extension de la première partie de l'algorithme (la détection des différences entre groupes dans les paramètres de seuil des items au premier temps de mesure), puisqu'il s'agit du point de départ avant la détection de la recalibration en prenant en compte plusieurs covariables.

Dans un cadre transversal, la première partie de ROSALI correspond à une recherche de DIF. Ainsi, afin de déterminer les différentes possibilités envisageables pour étendre la partie 1 de ROSALI, un travail de recherche bibliographique sur le DIF a été mené. L'objectif était d'identifier les méthodes de détection du DIF, basées sur l'IRT ou la RMT, permettant de

prendre simultanément en compte plusieurs covariables. Les méthodes que l'on a identifiées ont été présentées en détail dans le chapitre 3.

Nous avons proposé deux versions d'extension : les aglorithmes ROSALI-DIF FORWARD et ROSALI-DIF BACKWARD. Ces extensions conservent la philosophie de l'algorithme initial, mais s'inspirent également des travaux de Tay *et al.* [80, 89]. Les performances de ces deux extensions ont été évaluées en termes de taux de détection à raison et taux de détection à tort. Le taux de détection à tort correspondait à la proportion de réplications sur lesquelles du DIF était identifié alors qu'il n'était pas simulé. Le taux de détection à raison correspondait, quant à lui, à la proportion de réplications où le DIF simulé était retrouvé. Plusieurs taux de détection à raison ont été étudiés pour caractériser à quel point on retrouvait le DIF qui avait été simulé. En parallèle, nous avons évalué dans les mêmes conditions une méthode basée sur la pénalisation de la vraisemblance proposée par Schauberger et Mair [19]. Cette méthode a été choisie, car elle repose sur une philosophie très différente de celle de nos extensions (qui sont basées sur des tests, comme les autres méthodes identifiées dans la littérature).

En présence de deux covariables binaires, les extensions de la partie 1 de ROSALI (ROSALI-DIF FORWARD et BACKWARD) ont présenté des performances bonnes ou acceptables (tant que l'effectif et/ou la taille du DIF étaient suffisants). Si les deux extensions réussissaient globalement à retrouver les items affectés par du DIF et les covariables qui l'induisaient, elles présentaient néanmoins des difficultés à identifier la forme du DIF simulée. Les scénarios pour lesquels les deux extensions présentaient de mauvaises performances correspondaient à un DIF simulé avec une taille faible, qui n'entrainait généralement pas un biais substantiel dans l'estimation des effets des covariables (DIF non *"meaningful"*). Ces deux extensions présentaient également de meilleures performances que la méthode de pénalisation PCM-Lasso, aussi bien pour la détection à raison que la détection à tort. Pour rappel, cette procédure concluait à tort à la présence de DIF dans près de la moitié des réplications des scénarios sans DIF. Cela nous a par ailleurs permis d'avoir un regard critique sur les critères d'évaluation usuels des méthodes de détection du DIF (c.-à-d., les taux de vrais et faux positifs TPR et FPR).

Comme toute étude de simulation, cette étude présente des limites importantes. La principale étant que l'on ne connait les performances des procédures étudiées que dans les scénarios considérés. On dit que "les études de simulation éclairent certaines zones d'un paysage, mais ne peuvent pas éclairer l'ensemble" [161]. Avec notre étude de simulation, de nombreuses zones restent dans l'ombre et il serait intéressant de les explorer. Par ailleurs, des développements méthodologiques sont nécessaires pour considérer que l'effet DIF d'une covariable dépend du niveau de l'autre covariable (pour l'heure, les extensions proposées ne le permettent pas). Lorsque ce sera fait, il pourrait être intéressant de se confronter à la méthode MIMIC proposée par Chun *et al.* [22].

L'extension de la deuxième partie de ROSALI, dédiée à la détection de la recalibration reste à entreprendre. D'autres travaux sur cet algorithme sont également en cours au sein de SPHERE. Ils visent à : (i) introduire une covariables à trois modalités et (ii) prendre en compte plus de deux temps de mesure. On aimerait également étudier les performances de ROSALI en présence de données manquantes. Enfin, une implémentation sur le logiciel PRO-online (gratuit et accessible en ligne) permettra aux chercheurs d'utiliser l'algorithme ROSALI en se libérant des contraintes de codage. À terme, on souhaiterait donner la possibilité aux utilisateurs de choisir les items sur lesquels ils souhaitent rechercher du DIF et du *response shift*, et ceux qu'ils veulent considérer invariants. Cela permettrait d'intégrer des connaissances *a priori* dans l'analyse, afin de ne pas se baser uniquement sur des tests statistiques. Ces connaissances pourraient provenir de la littérature, ou d'entretiens qualitatifs [2].

Ces travaux sont également un premier pas en direction de la recherche de DIF causal (c'est-à-dire la recherche de DIF dont on peut effectivement affirmer qu'il est induit par la covariable à l'étude). Cette recherche de DIF causal peut être faite en ajustant l'analyse sur les facteurs de confusion potentiels. Néanmoins, le nombre de covariables que l'on peut introduire dans l'analyse avec un PCM risque d'être rapidement limité (le nombre de paramètres à estimer augmentant

2. Comme l'approche "pensée à voix haute" qui permet d'expliciter les processus cognitifs impliqués lors de la complétion d'un questionnaire et ainsi potentiellement identifier des items qui ne sont pas interprétés de la même façon entre différents groupes d'individus ou au cours du temps

très rapidement avec l'introduction de covariables interférant avec les paramètres de seuil des items). Une autre approche intéressante pour détecter du DIF causal a été proposé par Liu *et al.* [60]. La première étape de cette méthode vise à rendre comparables les groupes d'individus entre lesquels on recherche du DIF. Pour ce faire, ces auteurs utilisent un appariement sur score de propension. La détection du DIF est ensuite basée sur la procédure de la régression logistique [75]. Pour prendre en compte l'appariement des données, une régression logistique conditionnelle est utilisée (à la place d'une régression logistique traditionnelle). Si cette approche est intéressante, elle nécessite des développements méthodologiques pour être appliquée à la recherche de DIF causal au sein de données PRO. En effet, la méthode de détection du DIF par régression logistique est critiquée en présence d'un faible nombre d'items. À terme, on pourrait même envisager la détection de *response shift* causal, si l'on suspecte une caractéristique d'être à l'origine de la survenue de *response shift* dans le temps.

Propriétés psychométriques du PTGI

Dans la littérature, le *response shift* est présenté comme un phénomène pouvant être lié au développement post-traumatique. Plus précisément, la survenue de *response shift* pourrait être une conséquence d'un développement post-traumatique chez le répondant. Cette relation pourrait être intéressante pour étudier le *response shift* avec une granularité plus fine, car le développement post-traumatique peut être mesuré grâce à un questionnaire autorapporté (l'inventaire du développement post-traumatique, PTGI). Néanmoins, peu d'informations sont disponibles dans la littérature sur la validité de construit de ce questionnaire en langue française.

Le troisième chapitre de cette thèse s'est intéressé aux propriétés psychométriques d'une des versions françaises de ce questionnaire, chez des patients à qui on a diagnostiqué un cancer (sein ou mélanome) deux ans auparavant. Si cette étude n'informe pas complétement de la validité du questionnaire selon l'approche *argument-based*, elle permet de fournir des informations sur sa validité de construit et sa fiabilité. L'un des résultats principaux que l'on a obtenus est qu'une structure à quatre dimensions serait plus adéquate que la structure initiale à cinq dimensions

proposée par Tedeschi et Calhoun [24]. Au cours de cette étude, nous avons par ailleurs remarqué des changements substantiels entre la traduction étudiée et la formulation initiale des items et des modalités de réponse en anglais.

Malgré les limites évoquées, ce questionnaire nous semble être un outil intéressant à la fois pour la pratique clinique et pour la recherche (notamment dans l'objectif d'étudier les liens entre le développement post-traumatique et le *response shift*). En effet, comme nous l'avons évoqué, le *response shift* pourrait être une conséquence de la survenue de développement post-traumatique. Par exemple, lorsque l'on s'intéresse à l'évolution de la qualité de vie liée à la santé suite à un événement de vie défiant hautement les ressources des individus (comme l'annonce d'un cancer, ou l'initiation d'une dialyse), il est possible que les changements de priorités ou de directions de vie entraînent une dissonance entre l'évolution de la qualité de vie liée à la santé observée et celle qui était ciblée. Autrement dit, il est possible que le développement post-traumatique soit à l'origine du *response shift*.

Ces travaux visant à étudier les liens entre le *response shift* et le développement post-traumatique n'ont pas encore été initiés, mais on pourrait imaginer chercher du *response shift* parmi des individus présentant des profils de score spécifiques pour le PTGI. La temporalité des mesures qu'il serait pertinent d'utiliser pour ce type d'analyse reste néanmoins à déterminer. Il s'agit d'une question particulièrement importante : au bout de combien de temps pourrait-on discerner du *response shift* après qu'un patient ait perçu chez lui des changements psychologiques positifs ?

Deux études où le laboratoire SPHERE est impliqué pourraient notamment permettre de mener ce type d'analyse :

– L'étude ELCCA [193] où l'on dispose des mesures du développement post-traumatique et de la qualité de vie liée à la santé à 1 mois (HRQoL seulement), 6 mois, 1 an, 2 ans, 4 ans et 5 ans après le diagnostic d'un cancer du sein ou d'un mélanome. La qualité de vie liée à la santé est mesurée grâce au QLQ-C30 (questionnaire spécifique aux patients atteints d'un

cancer) [203].

– L'étude PrekitQoL où l'on dispose des mesures de développement post-traumatique et de la qualité de vie liée à la santé à 3 mois, 6 mois, 1 an, 2 ans, 3 ans, 4 ans et 5 ans après une greffe rénale. La qualité de vie liée à la santé est également mesurée avant la greffe, lorsque le patient est sur liste d'attente (la mesure s'effectue tous les 6 mois). Son évaluation est basée sur le SF-36 (questionnaire générique) [158] et le ReTransQoL [204] (questionnaire spécifique aux patients ayant eu une greffe rénale, adapté pour être également complété avant la greffe par les patients sur liste d'attente).

Le *response shift* que l'on pourra explorer grâce à ce type de travaux fera forcément suite à un événement "traumatisant". Il convient néanmoins de rappeler que les catalyseurs pouvant entraîner du *response shift* ne se limitent pas à ce type d'événements [58].

Épilogue

Pour pouvoir mieux analyser et interpréter les mesures issues des questionnaires autorapportés par les patients, il est capital d'impliquer les cliniciens. En France, ils sont relativement peu accoutumés aux problématiques que l'on peut rencontrer face à ce type de données. Il est donc important de les sensibiliser à ces problématiques pour que les données recueillies apportent un réel éclairage sur les décisions cliniques et viennent enrichir le dialogue patient-soignant (avec une compréhension mutuelle).

Par exemple, si l'on s'intéresse plus particulièrement au DIF et au *response shift*, il est important de faire comprendre aux cliniciens que l'interprétation des items d'un questionnaire peut parfois être très différente d'un individu à l'autre, et que cette interprétation des items n'est pas figée dans le temps. Kwon *et al.* [16] rappellent notamment que les cliniciens devraient utiliser les données rapportées par les patients comme un point de départ pour un dialogue avec le patient (ce qui permettrait ainsi d'éviter les erreurs d'interprétation qui risqueraient de brouiller la communication ou empêcher la prise de décision partagée). Par exemple, les cliniciens pourraient

chercher de plus amples informations auprès du patient lorsque les données qu'il/elle rapporte ne s'alignent pas avec son contexte clinique [16].

De fait, l'implication des cliniciens pourrait aider à améliorer la recherche sur le DIF et le *response shift*. Leurs observations pourraient permettre d'identifier des items et des caractéristiques des patients intéressants à investiguer. Les cliniciens pourraient également indiquer les types de *response shift* qui sont plutôt « positif » ou plutôt « négatif » pour le patient, mais également les cheminements qui conduisent les patients à ce type d'adaptation. Pour ce faire, il est nécessaire de sensibiliser les cliniciens au DIF et au *response shift*, grâce à des outils pédagogiques adaptés leur présentant ces phénomènes et leurs implications cliniques. À titre d'exemple, une équipe canadienne met actuellement en place une initiative dans ce sens. il s'agit d'une initiative interdisciplinaire de *knowledge translation* (transfert de connaissances) ciblant l'interprétation des données PRO par un public de cliniciens. Une partie de cette initiative s'intéresse plus particulièrement au DIF et au *response shift*, avec le développement d'images pédagogiques accompagnées d'exemples précis [16]. Ce type de ressources m'intéresse particulièrement et il serait intéressant d'adapter ces outils à un public de cliniciens français. De façon plus large, des réflexions sont en cours au sein du laboratoire SPHERE pour guider vers une meilleure utilisation des données rapportées par les patients et pour rechercher les façons les plus optimales de présenter les résultats extraits de ces données à différents types d'audience (chercheurs, cliniciens et patients).

Références

Références

[1] OMS. Constitution de l'Organisation Mondiale de la Santé, 1946. Disponible depuis https://www.who.int/fr/about/governance/constitution.

[2] FDA. Guidance for industry : patient-reported outcome measures : use in medical product development to support labeling claims : draft guidance. *Health and Quality of Life Outcomes*, 4(1) :79, 2006.

[3] Au, H.-J., Ringash, J., Brundage, M., Palmer, M., Richardson, H. et Meyer, R. M. Added value of health-related quality of life measurement in cancer clinical trials : the experience of the NCIC CTG. *Expert review of pharmacoeconomics & outcomes research*, 10(2) :119–128, 2010.

[4] Mercieca-Bebber, R., King, M. T., Calvert, M. J., Stockler, M. R. et Friedlander, M. The importance of patient-reported outcomes in clinical trials and strategies for future optimization. *Patient related outcome measures*, 9 :353, 2018.

[5] Bonnetain, F., Borg, C., Adams, R., Ajani, J., Benson, A., Bleiberg, H., Chibaudel, B., Diaz-Rubio, E., Douillard, J., Fuchs, C. *et al.* How health-related quality of life assessment should be used in advanced colorectal cancer clinical trials. *Annals of Oncology*, 28(9) :2077–2085, 2017.

[6] Fokkema, M., Smits, N., Kelderman, H. et Cuijpers, P. Response shifts in mental health interventions : an illustration of longitudinal measurement invariance. *Psychological assessment*, 25(2) :520, 2013.

[7] Johnston, B. C., Patrick, D. L., Devji, T., Maxwell, L. J., Bingham III, C. O., Beaton, D. E., Boers, M., Briel, M., Busse, J. W., Carrasco-Labra, A. *et al.* Patient-reported outcomes. *Cochrane handbook for systematic reviews of interventions*, pages 479–492, 2019.

[8] Verdam, M. G. E. *Using structural equation modeling to investigate change in health-related quality of life*. Manuscrit de thèse de doctorat, Université d'Amsterdam, 2017.

[9] Huber, M., Knottnerus, J. A., Green, L., Van Der Horst, H., Jadad, A. R., Kromhout, D., Leonard, B., Lorig, K., Loureiro, M. I., Van der Meer, J. W. *et al.* How should we define health ? *Bmj*, 343, 2011.

[10] Guillemin, F., Leplège, A., Briançon, S., Spitz, E. et Coste, J. *Perceived health and adaptation in chronic disease*. Routledge, 2017.

[11] L'Assurance Maladie. Améliorer la qualité du système de santé et maîtriser les dépenses, 2018. Disponible depuis https://www.ameli.fr/sites/default/files/rapport-activite-charges-produits-18_assurance-maladie.pdf.

[12] Fayers, P. M. et Machin, D. *Quality of life : the assessment, analysis and interpretation of patient-reported outcomes*. John Wiley & Sons, 2013.

[13] Osoba, D. Health-related quality of life and cancer clinical trials. *Therapeutic advances in medical oncology*, 3(2) :57–71, 2011.

[14] Ciani, O., Salcher-Konrad, M., Meregaglia, M., Smith, K., Gorst, S. L., Dodd, S., Williamson, P. R. et Fattore, G. Patient-reported outcome measures in core outcome sets targeted overlapping domains but through different instruments. *Journal of Clinical Epidemiology*, 136 :26–36, 2021.

[15] Mellenbergh, G. J. Item bias and item response theory. *International Journal of Educational Research*, 13(2) :127–143, January 1989.

[16] Kwon, J.-Y., Russell, L., Coles, T., Klaassen, R. J., Schick-Makaroff, K., Sibley, K. M., Mitchell, S. A. et Sawatzky, R. Patient-Reported Outcomes Measurement in Radiation Oncology : Interpretation of Individual Scores and Change over Time in Clinical Practice. *Current Oncology*, 29(5) :3093–3103, April 2022.

[17] Sprangers, M. A. et Schwartz, C. E. Integrating response shift into health-related quality of life research : a theoretical model. *Social Science & Medicine*, 48(11) :1507–1515, June 1999.

[18] Vanier, A., Oort, F. J., McClimans, L., Ow, N., Gulek, B. G., Böhnke, J. R., Sprangers, M., Sébille, V., Mayo, N. et Response Shift - in Sync Working Group. Response shift in patient-reported outcomes : definition, theory, and a revised model. *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, 30(12) :3309–3322, December 2021.

[19] Schauberger, G. et Mair, P. A regularization approach for the detection of differential item functioning in generalized partial credit models. *Behavior Research Methods*, 52(1) :279–294, February 2020.

[20] Tay, L., Vermunt, J. K. et Wang, C. Assessing the Item Response Theory With Covariate (IRT-C) Procedure for Ascertaining Differential Item Functioning. *International Journal of Testing*, 13(3) :201–222, July 2013.

[21] Jones, R. N. Differential item functioning and its relevance to epidemiology. *Current epidemiology reports*, 6 :174–183, June 2019.

[22] Chun, S., Stark, S., Kim, E. S. et Chernyshenko, O. S. MIMIC methods for detecting dif among multiple groups : Exploring a new sequential-free baseline procedure. *Applied psychological measurement*, 40(7) :486–499, 2016.

[23] Sébille, V., Lix, L. M., Ayilara, O. F., Sajobi, T. T., Janssens, A. C. J. W., Sawatzky, R., Sprangers, M. A. G., Verdam, M. G. E. et the Response Shift – in Sync Working Group. Critical examination of current response shift methods and proposal for advancing new methods. *Quality of Life Research*, 30(12) :3325–3342, December 2021.

[24] Tedeschi, R. G. et Calhoun, L. G. The Posttraumatic Growth Inventory : Measuring the positive legacy of trauma. *Journal of traumatic stress*, 9(3) :455–471, 1996.

[25] VandenBos, G. R. *APA dictionary of psychology.* American Psychological Association, 2007.

[26] Joreskog, K. G. et Van Thillo, M. Lisrel : A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables. 1972.

[27] Jöreskog, K. G. Structural analysis of covariance and correlation matrices. *Psychometrika*, 43(4) :443–477, 1978.

[28] Raykov, T. et Marcoulides, G. A. *A first course in structural equation modeling.* routledge, 2012.

[29] Satorra, A. et Bentler, P. M. Corrections to test statistics and standard errors in covariance structure analysis. 1994.

[30] Muthén, B. O., du Toit, S. H. C. et Spisic, D. Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical outcomes, 1997.

[31] Bandalos, D. L. Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling : a multidisciplinary journal*, 21(1) :102–116, 2014.

[32] Kline, R. B. *Principles and practice of structural equation modeling, 4th ed.* Guilford Press, New York, NY, US, 2016.

[33] Steiger, J. H. Statistically based tests for the number of common factors. In *the annual meeting of the Psychometric Society. Iowa City, IA. 1980*, 1980.

[34] Steiger, J. H. Structural model evaluation and modification : An interval estimation approach. *Multivariate behavioral research*, 25(2) :173–180, 1990.

[35] Browne, V. et Cudeck, R. Alternative Ways of Assessing Model Fit. In *Testing Structural Equation Models*. Newbury Park, CA : Sage, 1993.

[36] Bentler, P. M. *EQS structural equations program manual*, volume 6. Multivariate software Encino, CA, 1995.

[37] Bentler, P. M. Comparative fit indexes in structural models. *Psychological bulletin*, 107(2) :238, 1990.

[38] Hoyle, R. H. *Handbook of structural equation modeling*. Guilford press, 2012.

[39] Guttman, L. The basis for scalogram analysis. In *Measurement and prediction*. Princeton University Press, 1950.

[40] Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Studies in mathematical psychology. Danmarks Paedagogiske Institut, 1960.

[41] Birnbaum, A. L. Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*, 1968.

[42] Lord, F., Novick, M. et Birnbaum, A. *Statistical theories of mental test scores*. Statistical theories of mental test scores. Addison-Wesley, Oxford, England, 1968.

[43] Andrich, D. Controversy and the Rasch model : a characteristic of incompatible paradigms ? *Medical care*, pages I7–I16, 2004.

[44] Andrich, D. Rating scales and Rasch measurement. *Expert review of pharmacoeconomics & outcomes research*, 11(5) :571–585, 2011.

[45] Mousavi, A., Tendeiro, J. N. et Younesi, J. Person fit assessment using the PerFit package in R. *Quantitative methods for psychology*, 12(3) :232–242, 2016.

[46] Masters, G. N. A rasch model for partial credit scoring. *Psychometrika*, 47(2) :149–174, 1982.

[47] Andrich, D. A rating formulation for ordered response categories. *Psychometrika*, 43(4) :561–573, 1978.

[48] Andrich, D. Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied psychological measurement*, 2(4) :581–594, 1978.

[49] Andersen, E. B. Sufficient statistics and latent trait models. *Psychometrika*, 42(1) :69–81, 1977.

[50] Rasch, G. On specific objectivity an attempt at formalizing the request for generality and validity of scientific statements. *Danish yearbook of philosophy*, 14(1) :58–94, 1977.

[51] Little, R. J. et Rubin, D. B. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[52] Bock, R. D. et Aitkin, M. Marginal maximum likelihood estimation of item parameters : Application of an EM algorithm. *Psychometrika*, 46(4) :443–459, 1981.

[53] Muraki, E. A Generalized Partial Credit Model : Application of an EM Algorithm. *Applied Psychological Measurement*, 16(2) :159–176, June 1992.

[54] Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Eimicke, J. P., Crane, P. K., Jones, R. N., Lai, J.-S., Choi, S. W., Hays, R. D., Reeve, B. B., Reise, S. P., Pilkonis, P. A. et Cella, D. Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS) : An item response theory approach. *Psychology Science Quarterly*, 51(2) :148–180, 2009.

[55] Cole, S. R., Kawachi, I., Maller, S. J. et Berkman, L. F. Test of item-response bias in the CES-D scale. experience from the New Haven EPESE study. *Journal of Clinical Epidemiology*, 53(3) :285–289, March 2000.

[56] Gelin, M. N. et Zumbo, B. D. Differential Item Functioning Results May Change Depending On How An Item Is Scored : An Illustration With The Center For Epidemiologic Studies

Depression Scale. *Educational and Psychological Measurement*, 63(1) :65–74, February 2003.

[57] Bulteau, S. *Et si la dépression était un trouble de la (ré)évaluation de soi ? : intérêt du Response Shift dans l'analyse de l'effet des traitements sur l'évolution de la symptomatologie dépressive du point de vue des patients.* Manuscrit de thèse de doctorat, Université de Nantes, 2021.

[58] Sawatzky, R., Kwon, J.-Y., Barclay, R., Chauhan, C., Frank, L., van den Hout, W. B., Nielsen, L. K., Nolte, S., Sprangers, M. A. G. et the Response Shift – in Sync Working Group. Implications of response shift for micro-, meso-, and macro-level healthcare decision-making using results of patient-reported outcome measures. *Quality of Life Research*, 30(12) :3343–3357, December 2021.

[59] Zumbo, B. D. Three generations of DIF analyses : Considering where it has been, where it is now, and where it is going. *Language assessment quarterly*, 4(2) :223–233, 2007.

[60] Liu, Y., Zumbo, B., Gustafson, P., Huang, Y., Kroc, E. et Wu, A. Investigating causal DIF via propensity score methods. *Practical Assessment, Research, and Evaluation*, 21(1) :13, 2016.

[61] Sprangers, M. A. G. et Schwartz, C. E. Do not throw out the baby with the bath water : build on current approaches to realize conceptual clarity. Response to Ubel, Peeters, and Smith. *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, 19(4) :477–479, May 2010.

[62] Mayo, N. E. Appraisal as a unifying theory of response shift : Continuing the conversation. *Quality of Life Research*, 28(10) :2635–2636, 2019.

[63] Holland, P. W. et Thayer, D. T. Differential item functioning and the Mantel-Haenszel procedure. *ETS Research Report Series*, 1986(2) :i–24, 1986.

[64] Lord, F. M. *Applications of Item Response Theory To Practical Testing Problems*. Routledge, New York, July 1980.

[65] Mellenbergh, G. J. Applicability of the Rasch model in two cultures. *Mental tests and cultural adaptation*, pages 453–457, 1972.

[66] Mellenbergh, G. J. Contingency table models for assessing item bias. *Journal of educational statistics*, 7(2) :105–118, 1982.

[67] Hanson, B. A. Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics*, 23(3) :244–253, 1998.

[68] Shealy, R. et Stout, W. An item response model for test bias and differential item functioning. *Differential item functioning. Hillsdale, NJ : Lawrence Erlbaum*, 1993.

[69] Uttaro, T. et Millsap, R. E. Factors influencing the Mantel-Haenszel procedure the detection of differential item functioning. *Applied Psychological Measurement*, 18(1) :15–25, 1994.

[70] Millsap, R. E. *Statistical approaches to measurement invariance*. Routledge, 2012.

[71] Bollmann, S., Berger, M. et Tutz, G. Item-focused trees for the detection of differential item functioning in partial credit models. *Educational and Psychological Measurement*, 78(5) :781–804, 2018.

[72] Penfield, R. D. Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement*, 44(3) :187–210, 2007.

[73] Teresi, J. A. et Fleishman, J. A. Differential item functioning and health assessment. *Quality of Life Research*, 16(1) :33–42, 2007.

[74] Holland, P. W. et Thayer, D. T. Differential item performance and the Mantel-Haenszel procedure. In *Test validity*, pages 129–145. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, 1988.

[75] Swaminathan, H. et Rogers, H. J. Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, 27(4) :361–370, 1990.

[76] Millsap, R. E. et Everson, H. T. Methodology review : Statistical approaches for assessing measurement bias. *Applied psychological measurement*, 17(4) :297–334, 1993.

[77] Teresi, J. A. et Jones, R. N. Bias in psychological assessment and other measures. 2013.

[78] Tay, L., Meade, A. W. et Cao, M. An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, 18(1) :3–46, 2015.

[79] Magis, D., Béland, S., Tuerlinckx, F. et De Boeck, P. A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior research methods*, 42(3) :847–862, 2010.

[80] Tay, L., Newman, D. A. et Vermunt, J. K. Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence. *Organizational Research Methods*, 14(1) :147–176, 2011.

[81] Schwarz, G. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[82] Bozdogan, H. Model selection and Akaike's information criterion (AIC) : The general theory and its analytical extensions. *Psychometrika*, 52(3) :345–370, 1987.

[83] Bozdogan, H. Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix. In *Information and classification*, pages 40–54. Springer, 1993.

[84] Tay, L., Huang, Q. et Vermunt, J. K. Item response theory with covariates (IRT-C) assessing item recovery and differential item functioning for the three-parameter logistic model. *Educational and Psychological Measurement*, 76(1) :22–42, 2016.

[85] Muthen, B. Some uses of structural equation modeling in validity studies : Extending irt to external variables. *Test validity*, pages 213–238, 1988.

[86] Muthén, B. O. Latent variable modeling in heterogeneous populations. *Psychometrika*, 54(4) :557–585, 1989.

[87] Oort, F. J. Using restricted factor analysis to detect item bias. *Methodika*, 1992.

[88] Oort, F. J. Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling : A Multidisciplinary Journal*, 5(2) :107–124, 1998.

[89] Tay, L., Meade, A. W. et Cao, M. An Overview and Practical Guide to IRT Measurement Equivalence Analysis. *Organizational Research Methods*, 18(1) :3–46, January 2015.

[90] Oberski, D. L., van Kollenburg, G. H. et Vermunt, J. K. A Monte Carlo evaluation of three methods to detect local dependence in binary data latent class models. *Advances in Data Analysis and Classification*, 7(3) :267–279, 2013.

[91] Muthén, B. A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of educational statistics*, 10(2) :121–132, 1985.

[92] Muthen, B. Some Uses of Structural Equation Modeling in Validity Studies : Extending IRT to External Variables Using SIMS Results. Research on Instructional Assessment : Instructionally Relevant Psychometrics. 1986.

[93] Woods, C. M. Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, 44(1) :1–27, 2009.

[94] Woods, C. M. et Grimm, K. J. Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied psychological measurement*, 35(5) :339–361, 2011.

[95] Woods, C. M., Cai, L. et Wang, M. The Langer-improved Wald test for DIF testing with multiple groups : Evaluation and comparison to two-group irt. *Educational and Psychological Measurement*, 73(3) :532–547, 2013.

[96] Woods, C. M., Oltmanns, T. F. et Turkheimer, E. Illustration of MIMIC-model dif testing with the schedule for nonadaptive and adaptive personality. *Journal of psychopathology and behavioral assessment*, 31(4) :320–330, 2009.

[97] Wang, W.-C., Shih, C.-L. et Yang, C.-C. The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement*, 69(5) :713–731, 2009.

[98] Cheng, Y., Shao, C. et Lathrop, Q. N. The mediated MIMIC model for understanding the underlying mechanism of DIF. *Educational and Psychological Measurement*, 76(1) :43–63, 2016.

[99] Kim, E. S., Yoon, M. et Lee, T. Testing measurement invariance using MIMIC : Likelihood ratio test with a critical value adjustment. *Educational and Psychological Measurement*, 72(3) :469–492, 2012.

[100] Tutz, G. et Schauberger, G. A penalty approach to differential item functioning in Rasch models. *Psychometrika*, 80(1) :21–43, 2015.

[101] Tutz, G. et Berger, M. Item-focussed trees for the identification of items in differential item functioning. *Psychometrika*, 81(3) :727–750, 2016.

[102] Howard, G. S. et Dailey, P. R. Response-shift bias : A source of contamination of self-report measures. *Journal of Applied Psychology*, 64(2) :144–150, 1979.

[103] Golembiewski, R. T., Billingsley, K. et Yeager, S. Measuring Change and Persistence in Human Affairs : Types of Change Generated by OD Designs. *The Journal of Applied Behavioral Science*, 12(2) :133–157, April 1976.

[104] Breetvelt, I. et Van Dam, F. Underreporting by cancer patients : The case of response-shift. *Social Science & Medicine*, 32(9) :981–987, January 1991.

[105] Helson, H. *Adaptation-level theory : an experimental and systematic approach to behavior.*

Adaptation-level theory : an experimental and systematic approach to behavior. New York, Harper and Row, 1964.

[106] Lazarus, R. S. et Folkman, S. *Stress, appraisal, and coping*. Springer publishing company, 1984.

[107] Mishel, M. H. Uncertainty in illness. *Image : The Journal of Nursing Scholarship*, 20(4) :225–232, 1988.

[108] Rapkin, B. D. et Schwartz, C. E. Toward a theoretical model of quality-of-life appraisal : Implications of findings from studies of response shift. *Health and Quality of Life Outcomes*, 2 :14, March 2004.

[109] Tourangeau, R., Rips, L. J. et Rasinski, K., editors. *The psychology of survey response*. The psychology of survey response. Cambridge University Press, New York, NY, US, 2000. Pages : xiii, 401.

[110] Oort, F. J. Towards a formal definition of response shift (in reply to GW Donaldson). *Quality of Life Research*, pages 2353–2355, 2005.

[111] Oort, F. J., Visser, M. R. et Sprangers, M. A. Formal definitions of measurement bias and explanation bias clarify measurement and conceptual perspectives on response shift. *Journal of clinical epidemiology*, 62(11) :1126–1137, 2009.

[112] Schwartz, C. E. et Sprangers, M. A. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Social Science & Medicine (1982)*, 48(11) :1531–1548, June 1999.

[113] Barclay-Goddard, R., Epstein, J. D. et Mayo, N. E. Response shift : a brief overview and proposed research priorities. *Quality of Life Research*, 18(3) :335–346, 2009.

[114] Sajobi, T. T., Brahmbatt, R., Lix, L. M., Zumbo, B. D. et Sawatzky, R. Scoping review of response shift methods : current reporting practices and recommendations. *Quality of Life Research*, 27(5) :1133–1146, 2018.

[115] Sawatzky, R., Sajobi, T. T., Brahmbhatt, R., Chan, E. K., Lix, L. M. et Zumbo, B. D. Longitudinal change in response processes : A response shift perspective. In *Understanding and investigating response processes in validation research*, pages 251–276. Springer, 2017.

[116] Sprangers, M. A., Van Dam, F. S., Broersen, J., Lodder, L., Wever, L., Visser, M. R., Oosterveld, P. et Smets, E. M. Revealing response shift in longitudinal research on fatigue–the use of the thentest approach. *Acta Oncologica (Stockholm, Sweden)*, 38(6) :709–718, 1999.

[117] Nolte, S., Elsworth, G. R., Sinclair, A. J. et Osborne, R. H. Tests of measurement invariance failed to support the application of the "then-test". *Journal of clinical epidemiology*, 62(11) :1173–1180, 2009.

[118] Schwartz, C. E., Sprangers, M. A., Carey, A. et Reed, G. Exploring response shift in longitudinal data. *Psychology & Health*, 19(1) :51–69, 2004.

[119] Ruta, D. A., Garratt, A. M., Leng, M., Russell, I. T. et MacDonald, L. M. A new approach to the measurement of quality of life : the Patient-Generated Index. *Medical care*, pages 1109–1126, 1994.

[120] McGee, H. M., O'Boyle, C. A., Hickey, A., O'Malley, K. et Joyce, C. Assessing the quality of life of the individual : the SEIQoL with a healthy and a gastroenterology unit population. *Psychological medicine*, 21(3) :749–759, 1991.

[121] De Achaval, S., Kallen, M. A., Mayes, M. D., Lopez-Olivo, M. A. et Suarez-Almazor, M. E. Use of the patient-generated index in systemic sclerosis to assess patient-centered outcomes. *The Journal of rheumatology*, 40(8) :1337–1343, 2013.

[122] Ahmed, S., Mayo, N. E., Wood-Dauphinee, S., Hanley, J. A. et Cohen, S. R. Using the patient generated index to evaluate response shift post-stroke. *Quality of Life Research*, 14(10) :2247–2257, 2005.

[123] Ahmed, S., Mayo, N. E., Wood-Dauphinee, S., Hanley, J. A. et Cohen, S. R. The structural equation modeling technique did not show a response shift, contrary to the results of the then test and the individualized approaches. *Journal of Clinical Epidemiology*, 58(11) :1125–1133, 2005.

[124] Aburub, A., Gagnon, B., Ahmed, S., Rodríguez, A. et Mayo, N. E. Impact of reconceptualization response shift on rating of quality of life over time among people with advanced cancer. *Supportive Care in Cancer*, 26(9) :3063–3071, 2018.

[125] Botella, M., Zenasni, F., Pocard, M., Gledill, J. et Rodary, C. French adaptation of the patient generated index : metric characteristics and practical limitations. *Psycho-Oncologie*, 2 :131–140, 2007.

[126] Ring, L., Höfer, S., Heuston, F., Harris, D. et O'Boyle, C. A. Response shift masks the treatment impact on patient reported outcomes (PROs) : the example of individual quality of life in edentulous patients. *Health and quality of life outcomes*, 3(1) :1–8, 2005.

[127] O'Boyle, C. A., McGee, H. M. et Browne, J. P. Measuring response shift using the schedule for evaluation of individual quality of life. 2000.

[128] Korfage, I. J., de Koning, H. J. et Essink-Bot, M.-L. Response shift due to diagnosis and primary treatment of localized prostate cancer : a then-test and a vignette study. *Quality of Life Research*, 16(10) :1627–1634, 2007.

[129] Hinz, A., Häuser, W., Glaesmer, H. et Brähler, E. The relationship between perceived own health state and health assessments of anchoring vignettes. *International Journal of Clinical and Health Psychology*, 16(2) :128–136, 2016.

[130] Hinz, A. Using anchoring vignettes in the evaluation of breast cancer survivors' quality of life. *Breast Care*, 12(1) :33–37, 2017.

[131] Hinz, A., Karoff, J., Kittel, J., Brähler, E., Zenger, M., Schmalbach, B. et Kocalevent, R.-D. Associations between self-rated health and the assessments of anchoring vignettes in cardio-

vascular patients. *International Journal of Clinical and Health Psychology*, 20(2) :100–107, 2020.

[132] Preiß, M., Friedrich, M., Stolzenburg, J.-U., Zenger, M. et Hinz, A. Response shift effects in the assessment of urologic cancer patients' quality of life. *European Journal of Cancer Care*, 28(4) :e13027, 2019.

[133] Ware Jr, J. E., Kosinski, M. et Keller, S. D. A 12-Item Short-Form Health Survey : construction of scales and preliminary tests of reliability and validity. *Medical care*, pages 220–233, 1996.

[134] Topp, J., Heesen, C., Augustin, M., Andrees, V. et Blome, C. Challenges and lessons learned from using anchoring vignettes to explore quality of life response behavior. *Quality of Life Research*, 29(8) :2149–2159, 2020.

[135] Beeken, R. J., Eiser, C. et Dalley, C. Health-related quality of life in haematopoietic stem cell transplant survivors : a qualitative study on the role of psychosocial variables and response shifts. *Quality of life research*, 20(2) :153–160, 2011.

[136] Rapkin, B. D., Garcia, I., Michael, W., Zhang, J. et Schwartz, C. E. Distinguishing appraisal and personality influences on quality of life in chronic illness : Introducing the quality-of-life appraisal profile version 2. *Quality of Life Research*, 26(10) :2815–2829, 2017.

[137] Rapkin, B. D., Garcia, I., Michael, W., Zhang, J. et Schwartz, C. E. Development of a practical outcome measure to account for individual differences in quality-of-life appraisal : The brief appraisal inventory. *Quality of Life Research*, 27(3) :823–833, 2018.

[138] Verdam, M. et Oort, F. Conceptual and methodological considerations regarding appraisal and response shift. *Quality of Life Research*, 28(10) :2637–2639, 2019.

[139] Oort, F. J. Using structural equation modeling to detect response shifts and true change. *Quality of Life Research*, 14(3) :587–598, 2005.

[140] Verdam, M. G. et Oort, F. J. Measurement bias detection with kronecker product restricted models for multivariate longitudinal data : an illustration with health-related quality of life data from thirteen measurement occasions. *Frontiers in psychology*, 5 :1022, 2014.

[141] Verdam, M. et Oort, F. The analysis of multivariate longitudinal data : An instructive application of the longitudinal three-mode model. *Multivariate Behavioral Research*, 54(4) :457–474, 2019.

[142] King-Kallimanis, B., Oort, F. et Garst, G. Using structural equation modelling to detect measurement bias and response shift in longitudinal data. *AStA Advances in Statistical Analysis*, 94(2) :139–156, 2010.

[143] Lix, L. M., Chan, E. K., Sawatzky, R., Sajobi, T. T., Liu, J., Hopman, W. et Mayo, N. Response shift and disease activity in inflammatory bowel disease. *Quality of Life Research*, 25(7) :1751–1760, 2016.

[144] Verdam, M. G., Oort, F. J. et Sprangers, M. A. Using structural equation modeling to detect response shifts and true change in discrete variables : An application to the items of the SF-36. *Quality of Life Research*, 25(6) :1361–1383, 2016.

[145] Vanier, A., Sébille, V., Blanchin, M., Guilleux, A. et Hardouin, J.-B. Overall performance of Oort's procedure for response shift detection at item level : A pilot simulation study. *Quality of Life Research*, 24(8) :1799–1807, 2015.

[146] Schwartz, C. E. Introduction to special section on response shift at the item level. *Quality of Life Research*, 25(6) :1323–1325, June 2016.

[147] Guilleux, A., Blanchin, M., Vanier, A., Guillemin, F., Falissard, B., Schwartz, C. E., Hardouin, J.-B. et Sébille, V. RespOnse Shift ALgorithm in Item response theory (ROSALI) for response shift detection with missing data in longitudinal patient-reported outcome studies. *Quality of Life Research*, 24(3) :553–564, 2015.

[148] Blanchin, M., Sébille, V., Guilleux, A. et Hardouin, J.-B. The Guttman errors as a tool for response shift detection at subgroup and item levels. *Quality of Life Research*, 25(6) :1385–1393, June 2016.

[149] Blanchin, M., Guilleux, A., Hardouin, J.-B. et Sébille, V. Comparison of structural equation modelling, item response theory and Rasch measurement theory-based methods for response shift detection at item level : A simulation study. *Statistical Methods in Medical Research*, 29(4) :1015–1029, April 2020. Publisher : SAGE Publications Ltd STM.

[150] Hammas, K., Sébille, V., Brisson, P., Hardouin, J.-B. et Blanchin, M. How to Investigate the Effects of Groups on Changes in Longitudinal Patient-Reported Outcomes and Response Shift Using Rasch Models. *Frontiers in Psychology*, 11, 2020.

[151] Mayo, N. E., Scott, S. C., Dendukuri, N., Ahmed, S. et Wood-Dauphinee, S. Identifying response shift statistically at the individual level. *Quality of Life Research*, 17(4) :627–639, 2008.

[152] Akaike, H. Factor analysis and AIC. In *Selected papers of hirotugu akaike*, pages 371–386. Springer, 1987.

[153] Salmon, M., Blanchin, M., Rotonda, C., Guillemin, F. et Sébille, V. Identifying patterns of adaptation in breast cancer patients with cancer-related fatigue using response shift analyses at subgroup level. *Cancer Medicine*, 6(11) :2562–2575, 2017.

[154] Lix, L. M., Sajobi, T. T., Sawatzky, R., Liu, J., Mayo, N. E., Huang, Y., Graff, L. A., Walker, J. R., Ediger, J., Clara, I. *et al.* Relative importance measures for reprioritization response shift. *Quality of Life Research*, 22(4) :695–703, 2013.

[155] Li, Y. et Rapkin, B. Classification and regression tree uncovered hierarchy of psychosocial determinants underlying quality-of-life response shift in HIV/AIDS. *Journal of Clinical Epidemiology*, 62(11) :1138–1147, 2009.

[156] Boucekine, M., Boyer, L., Baumstarck, K., Millier, A., Ghattas, B., Auquier, P. et Toumi, M. Exploring the response shift effect on the quality of life of patients with schizophrenia : An application of the random forest method. *Medical Decision Making*, 35(3) :388–397, 2015.

[157] Sawatzky, R. Relating response shift and cognitive appraisal to measurement validation. *Quality of Life Research*, 28(10) :2633–2634, October 2019.

[158] Leplège, A., Ecosse, E., Verdier, A. et Perneger, T. V. The French SF-36 Health Survey : translation, cultural adaptation and preliminary psychometric evaluation. *Journal of clinical epidemiology*, 51(11) :1013–1023, 1998.

[159] Sijtsma, K. et Molenaar, I. *Introduction to nonparametric item response theory*. Number 5 in Measurement methods for the Social Science. Sage, 2002. Pagination : 176.

[160] Emons, W. H. M. Nonparametric Person-Fit Analysis of Polytomous Item Scores. *Applied Psychological Measurement*, 32(3) :224–247, May 2008.

[161] Morris, T. P., White, I. R. et Crowther, M. J. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11) :2074–2102, May 2019.

[162] Dubuy, Y., Sébille, V., Grall-Bronnec, M., Challet-Bouju, G., Blanchin, M. et Hardouin, J.-B. Evaluation of the link between the Guttman errors and response shift at the individual level. *Quality of Life Research*, 31(1) :61–73, 2022.

[163] Hardouin, J.-B. SIMIRT : Stata module to process data generated by IRT models. Statistical Software Components, Boston College Department of Economics, May 2005. Available from https://ideas.repec.org/c/boc/bocode/s450402.html (revised 16 may 2013).

[164] Penta, M., Arnould, C. et Decruynaere, C. Chapitre 5 La vérification des critères d'une mesure objective. In *Développer et interpréter une échelle de mesure*, Pratiques psychologiques, pages 87–114. Mardaga, Wavre, 2005.

[165] 28th Annual Conference of the International Society for Quality of Life Research. *Quality of Life Research*, 30(1) :1–177, October 2021.

[166] Wu, X., Sawatzky, R., Hopman, W., Mayo, N., Sajobi, T. T., Liu, J., Prior, J., Papaioannou, A., Josse, R. G., Towheed, T., Davison, K. S. et Lix, L. M. Latent variable mixture models to test for differential item functioning : a population-based analysis. *Health and Quality of Life Outcomes*, 15(1) :102, January 2017.

[167] Sawatzky, R., Ratner, P. A., Kopec, J. A. et Zumbo, B. D. Latent variable mixture models : a promising approach for the validation of patient reported outcomes. *Quality of Life Research*, 21(4) :637–650, 2012.

[168] Oliveri, M. E., Ercikan, K. et Zumbo, B. Analysis of sources of latent class differential item functioning in international assessments. *International Journal of Testing*, 13(3) :272–293, 2013.

[169] Maij-de Meij, A. M., Kelderman, H. et van der Flier, H. Improvement in detection of differential item functioning using a mixture item response theory model. *Multivariate Behavioral Research*, 45(6) :975–999, 2010.

[170] Cohen, A. S. et Bolt, D. M. A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, 42(2) :133–148, 2005.

[171] Samuelsen, K. *Examining differential item functioning from a latent class perspective*. University of Maryland, College Park, 2005.

[172] De Ayala, R. J., Kim, S.-H., Stapleton, L. M. et Dayton, C. M. Differential item functioning : A mixture distribution conceptualization. *International Journal of Testing*, 2(3-4) :243–276, 2002.

[173] Chen, Y.-F. et Jiao, H. Exploring the utility of background and cognitive variables in explaining latent differential item functioning : An example of the PISA 2009 reading assessment. *Educational Assessment*, 19(2) :77–96, 2014.

[174] Lee Webb, M.-y., Cohen, A. S. et Schwanenflugel, P. J. Latent class analysis of differential item functioning on the Peabody Picture Vocabulary Test–III. *Educational and Psychological Measurement*, 68(2) :335–351, 2008.

[175] DeMars, C. E. et Lau, A. Differential Item Functioning Detection With Latent Classes : How Accurately Can We Detect Who Is Responding Differentially ? *Educational and Psychological Measurement*, 71(4) :597–616, August 2011.

[176] Sajobi, T. T., Lix, L. M., Russell, L., Schulz, D., Liu, J., Zumbo, B. D. et Sawatzky, R. Accuracy of mixture item response theory models for identifying sample heterogeneity in patient-reported outcomes : a simulation study. *Quality of Life Research*, pages 1–10, 2022.

[177] Lix, L. M., Wu, X., Hopman, W., Mayo, N., Sajobi, T. T., Liu, J., Prior, J. C., Papaioannou, A., Josse, R. G., Towheed, T. E., Davison, K. S. et Sawatzky, R. Differential Item Functioning in the SF-36 Physical Functioning and Mental Health Sub-Scales : A Population-Based Investigation in the Canadian Multicentre Osteoporosis Study. *PLOS ONE*, 11(3) :e0151519, March 2016. Publisher : Public Library of Science.

[178] Blanchin, M., Brisson, P. et Sébille, V. Performance of a Rasch-based method for group comparisons of longitudinal change and response shift at the item level in PRO data : A simulation study. *Methods*, page S1046202322000020, January 2022.

[179] Cao, M., Tay, L. et Liu, Y. A Monte Carlo Study of an Iterative Wald Test Procedure for DIF Analysis. *Educational and Psychological Measurement*, 77(1) :104–118, January 2017.

[180] Rouquette, A., Hardouin, J.-B. et Coste, J. Differential Item Functioning (DIF) and Subsequent Bias in Group Comparisons using a Composite Measurement Scale : A Simulation Study. *Journal of Applied Measurement*, 17(3) :312–334, 2016.

[181] Tutz, G. et Schauberger, G. A Penalty Approach to Differential Item Functioning in Rasch Models. *Psychometrika*, 80(1) :21–43, March 2015.

[182] Lelorain, S., Bonnaud-Antignac, A. et Florin, A. Long term posttraumatic growth after breast cancer : prevalence, predictors and relationships with psychological health. *Journal of clinical psychology in medical settings*, 17(1) :14–22, 2010.

[183] Lelorain, S. *Qualité de vie et développement post-traumatique à long terme d'un cancer du sein*. Manuscrit de thèse de doctorat, Université de Nantes, 2009.

[184] Tedeschi, R. G. et Calhoun, L. G. *Trauma and transformation*. Sage, 1995.

[185] Tedeschi, R. G. et Calhoun, L. G. " posttraumatic growth : conceptual foundations and empirical evidence". *Psychological inquiry*, 15(1) :1–18, 2004.

[186] Tedeschi, R. G., Shakespeare-Finch, J., Taku, K. et Calhoun, L. G. *Posttraumatic growth : Theory, research, and applications*. Routledge, 2018.

[187] Beck, J. S. *Cognitive behavior therapy : Basics and beyond*. Guilford Publications, 2020.

[188] Martin-Krumm, C. et Tarquinio, C. *Traité de psychologie positive*. De Boeck, 2011.

[189] De Ridder, D., Geenen, R., Kuijer, R. et van Middendorp, H. Psychological adjustment to chronic disease. *The Lancet*, 372(9634) :246–255, 2008.

[190] Garrido-Hernansaiz, H., Rodríguez-Rey, R., Collazo-Castiñeira, P. et Collado, S. The post-traumatic growth inventory-short form (PTGI-SF) : A psychometric study of the spanish population during the COVID-19 pandemic. *Current Psychology*, pages 1–10, 2022.

[191] Cadell, S., Suarez, E., Hemsworth, D. *et al.* Reliability and validity of a French version of the posttraumatic growth inventory. *Open Journal of Medical Psychology*, 4(02) :53, 2015.

[192] Dubuy, Y., Sébille, V., Bourdon, M., Hardouin, J.-B. et Blanchin, M. Posttraumatic growth inventory : challenges with its validation among French cancer patients. *BMC medical research methodology*, 22(1) :1–18, 2022.

[193] Bourdon, M., Blanchin, M., Tessier, P., Campone, M., Quéreux, G., Dravet, F., Sébille, V. et Bonnaud-Antignac, A. Changes in quality of life after a diagnosis of cancer : a

2-year study comparing breast cancer and melanoma patients. *Quality of Life Research*, 25(8) :1969–1979, 2016.

[194] Tedeschi, R. G., Cann, A., Taku, K., Senol-Durak, E. et Calhoun, L. G. The posttraumatic growth inventory : A revision integrating existential and spiritual change. *Journal of Traumatic Stress*, 30(1) :11–18, 2017.

[195] Porro, B., Broc, G., Baguet-Marin, F. et Cousson-Gélie, F. A questionable version of the Post-Traumatic growth inventory—Short form in women diagnosed with breast cancer. *British Journal of Health Psychology*, 2022.

[196] Evans, C., Saliba-Serre, B., Préau, M., Bendiane, M.-K., Gonçalves, A., Signoli, M. et Bouhnik, A.-D. Post-traumatic growth 5 years after cancer : identification of associated actionable factors. *Supportive Care in Cancer*, pages 1–10, 2022.

[197] Henson, C., Truchot, D., Canevello, A. et Andela, M. Psychometric Properties of a European French Version of the PTGI. *Research on Social Work Practice*, page 10497315221101906, 2022.

[198] Cann, A., Calhoun, L. G., Tedeschi, R. G., Taku, K., Vishnevsky, T., Triplett, K. N. et Danhauer, S. C. A short form of the Posttraumatic Growth Inventory. *Anxiety, Stress, & Coping*, 23(2) :127–137, 2010.

[199] Prati, G. et Pietrantoni, L. Italian adaptation and confirmatory factor analysis of the full and the short form of the posttraumatic growth inventory. *Journal of Loss and Trauma*, 19(1) :12–22, 2014.

[200] Kaur, N., Porter, B., LeardMann, C. A., Tobin, L. E., Lemus, H. et Luxton, D. D. Evaluation of a modified version of the Posttraumatic Growth Inventory-Short Form. *BMC Medical Research Methodology*, 17(1) :1–9, 2017.

[201] Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M. et De Vet, H. C. The COSMIN checklist for assessing the methodological

quality of studies on measurement properties of health status measurement instruments : an international Delphi study. *Quality of life research*, 19(4) :539–549, 2010.

[202] Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M. et de Vet, H. C. COSMIN checklist manual. *Amsterdam : University Medical Center*, 2012.

[203] Scott, N. W., Fayers, P., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., Gundy, C., Koller, M., Petersen, M. A., Sprangers, M. A. *et al.* EORTC QLQ-C30 reference values manual. 2008.

[204] Gentile, S., Jouve, E., Dussol, B., Moal, V., Berland, Y. et Sambuc, R. Development and validation of a French patient-based health-related quality of life instrument in kidney transplant : the ReTransQoL. *Health and quality of life outcomes*, 6(1) :1–12, 2008.

Annexes

Annexe A

Article :

Identification of sources of DIF using covariates : a simulation study comparing two approaches based on Rasch family models

Title: Identification of sources of DIF using covariates: a simulation study comparing two approaches based on Rasch family models

Authors: Yseulys Dubuy (Y.D)[1*], Jean-Benoit Hardouin (JB.H)[1,2,3], Myriam Blanchin (M.B)[1‡], Véronique Sébille (V.S)[1,2‡]

Affiliations:

[1] Nantes Université, Université de Tours, INSERM, MethodS in Patients-centered outcomes and HEalth Research, SPHERE, F-44000 Nantes, France
[2] Nantes Université, CHU Nantes, Methodology and Biostatistics Unit, F-44000 Nantes, France;
[4] Nantes Université, CHU Nantes, Public Health Department, F-44000 Nantes, France

‡These authors contributed equally

ORCiD: Yseulys Dubuy: 0000-0001-8390-2285, Véronique Sébille: 0000-0002-0780-7742, , Myriam Blanchin: 0000-0003-1318-7620, Jean-Benoit Hardouin: 0000-0001-8664-623X

***Corresponding author information:**

Yseulys Dubuy

UMR INSERM 1246 - SPHERE
Institut de recherche en Santé 2 (IRS2)
22 boulevard Benoni-Goullin
44200 Nantes
France

e-mail: yseulys.dubuy@univ-nantes.fr

Abstract: Methods to detect and account for differential item functioning (DIF) are numerous. However, most of them only enable the detection of DIF considering one covariate. Hence, when searching for DIF using multiple covariates, the analysis must be repeated for each covariate. One issue, among others, is that it may lead to the detection of false-positive effects when covariates are correlated. We extended the ROSALI algorithm dedicated to the assessment of measurement invariance to obtain an iterative item-by-item DIF detection method based on Rasch family models that enable to detect and adjust group comparisons for DIF. This algorithm was evaluated through a simulation study under various conditions aiming to be representative of the health research framework. The performance of the algorithm was assessed using the rates of false and correct detection of DIF, the DIF size and form recovery, and the bias in the latent variable level estimation. We compared the performance of ROSALI to the one of an approach based on likelihood penalization. The extension of ROSALI performed better overall than the penalized likelihood approach.

Keywords: Differential item functioning, DIF, Rasch models, Partial Credit Model, Regularization, Lasso

Introduction

Patient-reported outcome (PRO) measures have gained interest in health research to take into account patients' perspectives on healthcare (1). PRO measures are often obtained via questionnaires usually completed by patients. These questionnaires include several items usually grouped into one or several domains to measure unobservable constructs (i.e. *latent variables*) such as fatigue or anxiety. Studies involving PRO measures often aim to compare patient levels on a latent variable by means of group comparisons and/or to study change in the latent variable. To make valid comparisons, one must ensure that individuals with different characteristics interpret the items in the same way and/or that their perception of the items remains the same over time (2). However, patients' characteristics may interfere with how some items are perceived. This phenomenon is known as *differential item functioning* (DIF). DIF occurs when patients do not interpret items in the same way according to their group membership and thus have differing item endorsement probabilities despite having the same latent variable level. In case of DIF, there is a violation of the assumption of between-group measurement invariance (3–5). Ignoring this lack of measurement invariance may lead to measurement bias, as observed between-group differences may not only reflect differences in the targeted latent variable (6). Changes in the meaning of the subjective evaluation of the target construct may also occur over time, leading to noncomparable data between time points due to a lack of longitudinal measurement invariance. This phenomenon has been acknowledged as *response shift* (7,8).

There is a wide range of DIF detection methods in the literature. Among them, we can mention the Mantel-Haenszel method (9), the logistic regression procedure (10), the likelihood-ratio test (11–13), the Lord's chi-square (Wald) test (14). Most of the existing methods only allow considering one covariate at a time. Yet, sources of DIF can be multiple (15), and there may be more than one covariate of interest. For instance, perception of items might differ according to gender but also age or health status. Moreover, there may be situations where two correlated covariates are investigated for DIF, but only one is really inducing DIF. Applying one-covariate methods separately could lead to inferring DIF for the wrong covariate in addition to the DIF-covariate due to the correlation between the two. Hence, using such an approach might not be appropriate to disentangle DIF effects between several covariates. Therefore, DIF detection considering simultaneously several covariates, possibly correlated, could be of great interest to get more insight into the sources of DIF. Several statistical approaches based on item response theory (IRT) or Rasch measurement theory (RMT) have been recently developed for that purpose. On the one hand, we can mention iterative detection methods such as the IRT with covariates (IRT-C) procedure (16,17) and the recursive partitioning approaches (namely the partial credit model (PCM) tree, PCM-tree (18) and the item-focused tree algorithm, PCM-IFT (19)). Yet, these methods show

some limitations. Indeed, the IRT-C procedure is only designed for dichotomous items, and the indices on which the procedure relies have been questioned (20). Besides, the PCM-tree approach makes it hard to identify which item is affected by DIF (19) and the current implemented version of the PCM-IFT algorithm does not seem to model the covariates' effect on the latent variable level (adjusted for DIF when appropriate). On the other hand, Schauberger and Mair proposed two methods based on penalized estimation of IRT or RMT models: one that only searches for a specific form of DIF having the same effects across all response categories (evaluated by simulations) and one that searches for more general forms of DIF not assuming that DIF has the same effect across all response categories (not evaluated by simulations) (21). Data on the DIF detection performance of these penalization-based approaches in case of simultaneous covariates are lacking as simulations pertained to a specific form of DIF in polytomous items. In addition, simulated tests were always composed of 20 items, which is rarely the case in health research, where the domains of the most commonly used scales include between 2 and 10 items (e.g. SF-36, HADS or PROMIS-29 (22–24)).

In a broader issue of measurement invariance assessment, the ROSALI algorithm (25–27) has been proposed in the RMT framework to detect and adjust for DIF and response shift in the analysis of longitudinal PRO data (polytomous and dichotomous items) in order to ensure valid comparisons between groups and over time. ROSALI is an iterative item-by-item detection algorithm that currently enables the introduction of one binary covariate in the analysis. It consists of two main parts that allow to:

- Identify items that function differently between the two groups defined by the covariate at the first measurement occasion (first part of ROSALI).
- Determine whether the perception of some items changes between two time points and assess whether or not these changes over time are similar in both groups (second part of ROSALI).

Of note, ROSALI ends by a final model allowing to adjust latent variable levels comparisons for the lack of invariance previously evidenced, if appropriate. Simulations showed that ROSALI does not erroneously infer DIF when DIF has not been simulated (27) and its performance to detect DIF are currently being assessed with one covariate in another study. To date, there is a will to extend ROSALI to simultaneously consider several sources of lack of invariance (e.g., gender, country). Thus, the first part of ROSALI needs to be developed to detect and adjust for DIF in presence of several covariates. However, it is currently unclear whether item-by-item iterative processes are the best approach or if it would be better to use a penalization approach that allows searching for DIF in all items simultaneously.

The aim of this study is twofold:

(1) To extend the first part of ROSALI (dedicated to the detection of DIF) to enable the simultaneous introduction of two binary covariates

(2) To compare by simulations the detection performance of this extension to the one obtained with the approach using penalization of likelihood under various conditions, including moderate numbers of polytomous items (representative of PRO measures used in health research), potentially correlated covariates, and various forms of DIF.

Methods

Rasch measurement theory

RMT is a family of models derived from the Rasch model for dichotomous items (28). For polytomous items, the most flexible model is the PCM (29,30), its formulation for a questionnaire composed of J polytomous items with M_j response categories for item j $(j = 1, \dots, J)$ is given by:

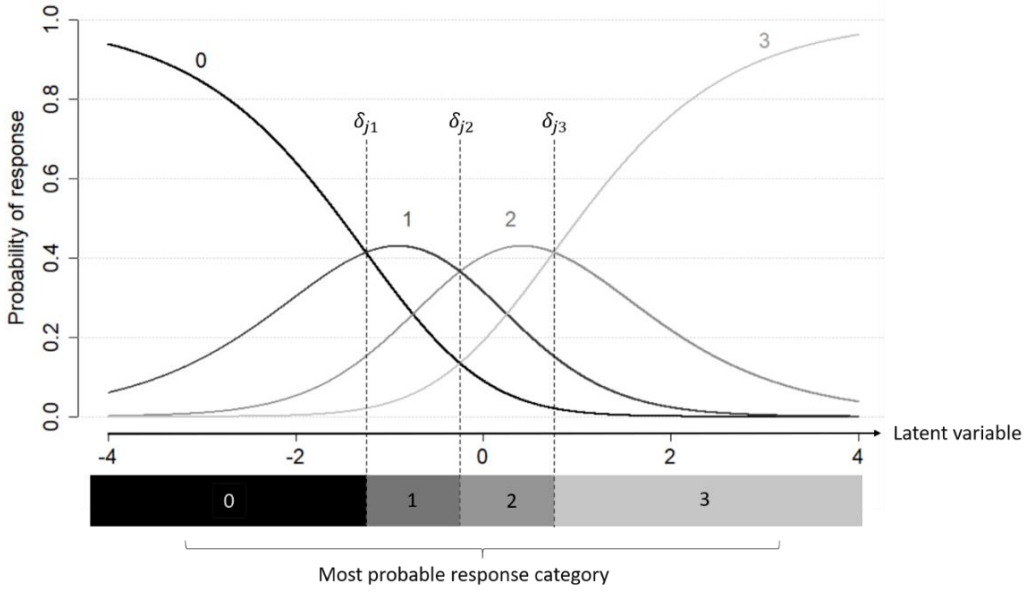
$$\mathbb{P}\left(X_{ij} = x \,\middle|\, \theta_i, \delta_{j1}, \dots, \delta_{jM_j-1}\right) = \frac{\exp\left(x\theta_i - \sum_{p=1}^{x} \delta_{jp}\right)}{\sum_{l=0}^{M_j-1} \exp\left(l\theta_i - \sum_{p=1}^{l} \delta_{jp}\right)}$$

The conditional probability that an individual i answers x $(= 0, 1, \dots, M_j - 1)$ to item j is a function of:

- The latent variable level of individual i: θ_i

  Where θ_i is the realization of Θ, a random variable assumed normally distributed (with mean μ and standard deviation σ). This latent variable is assumed to represent the target construct (e.g. anxiety).
- The *item threshold parameters* δ_{jp} associated with each response category $p > 0$ of item j $(1 \leq p \leq M_{j-1})$. δ_{jp} represents the latent variable level at which the probabilities of answering category p or $p - 1$ to item j are equal. When tracing the probability curves of each response category, item threshold parameters (e.g., δ_{j1}) correspond to the intersection between two adjacent category probability curves as pictured in Figure 1 (31).

Fig 1. Category characteristic curves for a given item j with four response categories under a partial credit model. Item threshold parameters δ_{jp} are indicated by dashed lines



DIF in RMT

Within RMT, DIF has been operationalized as between-group differences in the item threshold parameters. Dichotomous items are characterized by a single threshold parameter. Hence, DIF in dichotomous items can take on a single form, where category probability curves are shifted between groups (i.e., *parallel uniform DIF* (3)). However, polytomous items are characterized by several threshold parameters (one for each response category above 0). Therefore, between-group differences in the item threshold parameters may vary in magnitude, direction, or both, leading to numerous potential DIF forms (32). For instance, between-group differences in the item threshold parameters can:

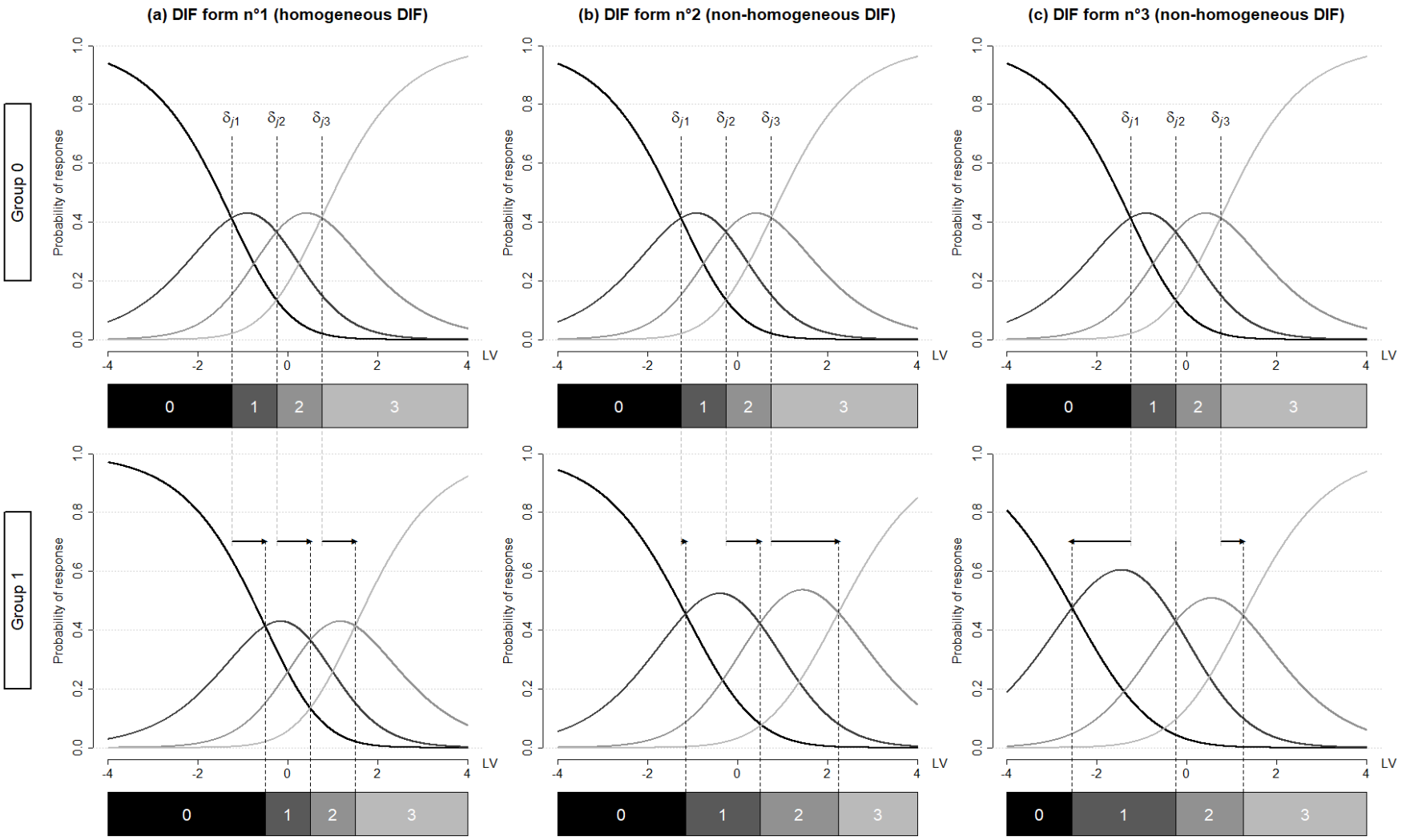
(1) Have the same direction and the same magnitude, e.g., Figure 2.a

(2) Have the same direction but vary in magnitude; e.g., Figure 2.b

(3) Vary in direction and magnitude, e.g., Figure 2.c

In the manuscript, the forms described in (2) and (3) will be referred to as *non-homogeneous DIF* (19). Of note, these two forms illustrate respectively the *convergent* and *divergent differential step functioning* introduced by Penfield et al. [32,33]. To maintain a consistent terminology throughout the manuscript, the form described in (1) will be referred to as *homogeneous DIF*. Of note, Penfield et al. referred to it to as *pervasive constant differential step functioning* and many researchers use the term DIF (e.g., [19,21]).

Fig 2. Category characteristic curves for a given item *j* affected by differential item functioning (DIF). DIF is operationalized by between-group differences in the item threshold parameters. These differences are represented by arrows. LV: Latent variable



The PCM can be used to assess the impact of a binary covariate C on the latent variable level accounting for a potential DIF induced by C through the introduction of group effects on the latent variable level and on the item threshold parameters:

$$\mathbb{P}\left(X_{ij} = x \left| \theta_i, C_i, \beta, \delta_{j1}, \dots, \delta_{jM_j-1}, \gamma_{j1}, \dots, \gamma_{jM_j-1} \right.\right) = \frac{\exp\left(x[\theta_i + \beta.C_i] - \sum_{p=1}^{x}[\delta_{jp} + \gamma_{jp}.C_i]\right)}{\sum_{l=0}^{M_j-1}\exp\left(l[\theta_i + \beta.C_i] - \sum_{p=1}^{l}[\delta_{jp} + \gamma_{jp}.C_i]\right)} \quad (1)$$

In addition to the above-mentioned parameters we have:

- C_i the realization of covariate C for individual i. C_i equals either 0 (reference group) or 1.
- β the effect of covariate C on the latent variable level (sometimes referred to as the *group effect*). β equals the difference between μ_1 and μ_0, where μ_1 designates the latent variable mean in the group of individuals with $C_i = 1$, and μ_0 designates the latent variable mean in the group $C_i = 0$ ($\beta = \mu_1 - \mu_0$)
- γ_{jp} the DIF parameters interfering with the item thresholds and modeling the DIF effects of covariate C. These DIF parameters operationalize the difference in item threshold parameters between the groups. Item threshold parameters in the reference group are δ_{jp} and item threshold parameters in the focal group are equal to $\delta_{jp} + \gamma_{jp}$. If there is no DIF on item j, then $\gamma_{jp} = 0$.

Additional binary covariates can be added in the same way. For instance, with two covariates C1 and C2, without interaction:

$$\mathbb{P}\left(X_{ij} = x \left| \theta_i, C_{1_i}, \beta_1, C_{2_i}, \beta_2, \delta_{j1}, \dots, \delta_{jM_j-1}, \gamma_{j1}^{(C_1)}, \dots, \gamma_{jM_j-1}^{(C_1)}, \gamma_{j1}^{(C_2)}, \dots, \gamma_{jM_j-1}^{(C_2)} \right.\right) =$$

$$\frac{\exp\left(x[\theta_i + \beta_1.C_{1_i} + \beta_2.C_{2_i}] - \sum_{p=1}^{x}[\delta_{jp} + \gamma_{jp}^{(C_1)}.C_{1_i} + \gamma_{jp}^{(C_2)}.C_{2_i}]\right)}{\sum_{l=0}^{M_j-1}\exp\left(l[\theta_i + \beta_1.C_{1_i} + \beta_2.C_{2_i}] - \sum_{p=1}^{l}[\delta_{jp} + \gamma_{jp}^{(C_1)}.C_{1_i} + \gamma_{jp}^{(C_2)}.C_{2_i}]\right)} \quad (2)$$

DIF detection procedures

Extension of the first part of ROSALI

The first part of the ROSALI algorithm with one binary covariate has been described elsewhere (25,27). We extended this algorithm by adding a second binary covariate. DIF detection then relies on the following steps:

Step 1. Estimation of a fully non-invariant PCM where the two covariates are assumed to induce DIF on all items.

Step 2. Estimation of a fully invariant PCM (no DIF is assumed).

Step 3. Test of the global occurrence of DIF by comparing the two previous models using a likelihood-ratio test (LRT).

Step 4. If the LRT is significant, screen all item-covariate pairs for DIF separately based on the fully non-invariant model. Otherwise, go to step 6.

Step 5. Forward iterative selection of the significant DIF item-covariate pairs found in step 4 (starting from the fully invariant model) and assessment of the form of DIF involved.

Step 6. Estimation of a final model giving the covariates effect on the latent variable level adjusted for DIF (if appropriate).

This extension of the first part of ROSALI will be referred to as ROSALI-DIF FORWARD. All steps are comprehensively described in Table 1 alongside statistical considerations. Of note, an alternative version of this algorithm has also been explored, with the same philosophy, but with an iterative step based on a backward instead of a forward process where all candidate pairs are tested simultaneously instead of one-by-one. This alternative version has been named ROSALI-DIF BACKWARD and is described in online appendix A. Both algorithms are jointly pictured in Figure 3. Of note, the screening step (step 4) was inspired by the iterative Wald test procedure [34,35].

Table 1: Comprehensive description of the ROSALI-DIF FORWARD algorithm and statistical considerations

ROSALI-DIF FORWARD steps	Statistical considerations
Step 1: Estimation of a fully non-invariant model (Model A) A fully unconstrained PCM is estimated in this first step where the two binary covariates C_1 and C_2 are assumed to induce DIF on all items	All DIF parameters $\gamma_{jp}^{(C_1)}$ and $\gamma_{jp}^{(C_2)}$ are freely estimated ($\forall\, j$ and p) in Equation 2. **Identifiability constraints:** the effects of covariates on the latent variable level are constrained to 0 ($\beta_1 = \beta_2 = 0$).
Step 2: Estimation of a fully invariant model (Model B) A fully constrained model assuming no DIF is estimated in this second step.	All DIF parameters $\gamma_{jp}^{(C_1)}$ and $\gamma_{jp}^{(C_2)}$ are constrained to zero ($\forall\, j$ and p) in Equation 2. The effects of covariates on the latent variable level (i.e., β_1 and β_2) are freely estimated.
Step 3. Test of the global occurrence of DIF The third step aims to evaluate the global occurrence of DIF by comparing model A and model B using a likelihood-ratio test. If the test is not significant, we assume that the covariates do not induce DIF and the algorithm moves directly to step 6 where the final model is model B. Otherwise, we proceed to the next step.	**Rationale for the likelihood-ratio test:** Model B is nested in Model A **Significance level:** 5%
Step 4. Screen item-covariate pairs (Item j, Covariate C) candidate for DIF detection From Model A (where the two covariates induce DIF on all items), statistical tests are performed for each item-covariate pair separately to determine whether the DIF effect induced by covariate C on item j is significant or not. Candidate pairs are those associated with significant tests. Measurement invariance is assumed for the other pairs (anchor pairs). Of note, if no pairs are considered as candidate, the algorithm goes directly to step 6 where the final model is model B.	**Statistical tests**: Contrast test **Null and alternative hypotheses of contrast test for DIF:** $H_0)\ \forall\, p,\ \gamma_{jp}^{(C)} = 0$ (No DIF) $H_1)\ \exists\, p:\ \gamma_{jp}^{(C)} \neq 0$ (DIF) **Significance level:** 5%
Step 5. Selection of DIF item-covariate pairs (Item j, Covariate C) among candidate pairs and assessment of the form of DIF involved This step is an iterative step that aims to select the item-covariate pairs affected by DIF among candidate pairs and determine the form of DIF involved. A new model (Model C) is introduced so that Model C = Model B at the beginning of this step. From model C, we estimate new models (one for each candidate pair) where the invariance constraint associated with the pair of interest is relaxed, and other constraints remain unchanged. From these new models, statistical tests are performed for each pair to determine whether the DIF effect induced by covariate C on item j is significant or not. We retain the model with the pair having the most significant test (smallest p-value) after Bonferroni correction. The associated pair is assumed to be affected by DIF and will be denoted (item j^*, covariate C^*). If there is no significant pair, the algorithm moves to step 6. Otherwise, based on the retained model, the form of DIF induced by covariate C^* on item j^* is assessed using another test. Model C is updated to account for the evidenced DIF and its form. The retained pair will no longer be tested. Step 5 is repeated over the remaining pairs to be tested. The step ends if no more pair is retained, if all candidate pairs have been tested, or just before relaxing the invariance constraint of the last anchor item for a given covariate.	********** Test DIF effect of candidate pairs ********** **Null and alternative hypotheses of contrast test for DIF:** $H_0)\ \forall\, p,\ \gamma_{jp}^{(C)} = 0$ (No DIF) $H_1)\ \exists\, p:\ \gamma_{jp}^{(C)} \neq 0$ (DIF) **Significance level**: 5%/number of candidate pairs, Bonferroni correction performed to avoid the inflation of the type I error rate due to multiple testing. ********** Test DIF form on the retained pair ********** **Null and alternative hypotheses of contrast test to assess DIF form:** $H_0)\ \forall\, p,\ \gamma_{j^*p}^{(C^*)} = \gamma_{j^*}^{(C^*)}$ (Homogeneous DIF) $H_1)\ \exists\, p,\ p':\ \gamma_{j^*p}^{(C^*)} \neq \gamma_{j^*p'}^{(C^*)}$ (Non-homogeneous DIF) **Significance level:** 5% ********** Update Model C ********** If the previous test is significant, the DIF parameters $\gamma_{j^*p}^{(C^*)}$ associated with the retained pair are freely estimated (non-homogeneous DIF). Otherwise, the DIF parameters $\gamma_{j^*p}^{(C^*)}$ are estimated but constrained to be constant over all response categories (homogeneous DIF).
Step 6. Estimation of the covariates effect on the latent variable level (Model D) The last step estimates the effect of the covariates C_1 and C_2 on the latent variable level adjusted for the DIF that was previously evidenced, if appropriate, using a final model called model D.	Model D = Model B if no DIF has been evidenced. Otherwise, model D is equal to the last version of model C obtained at the end of step 5.

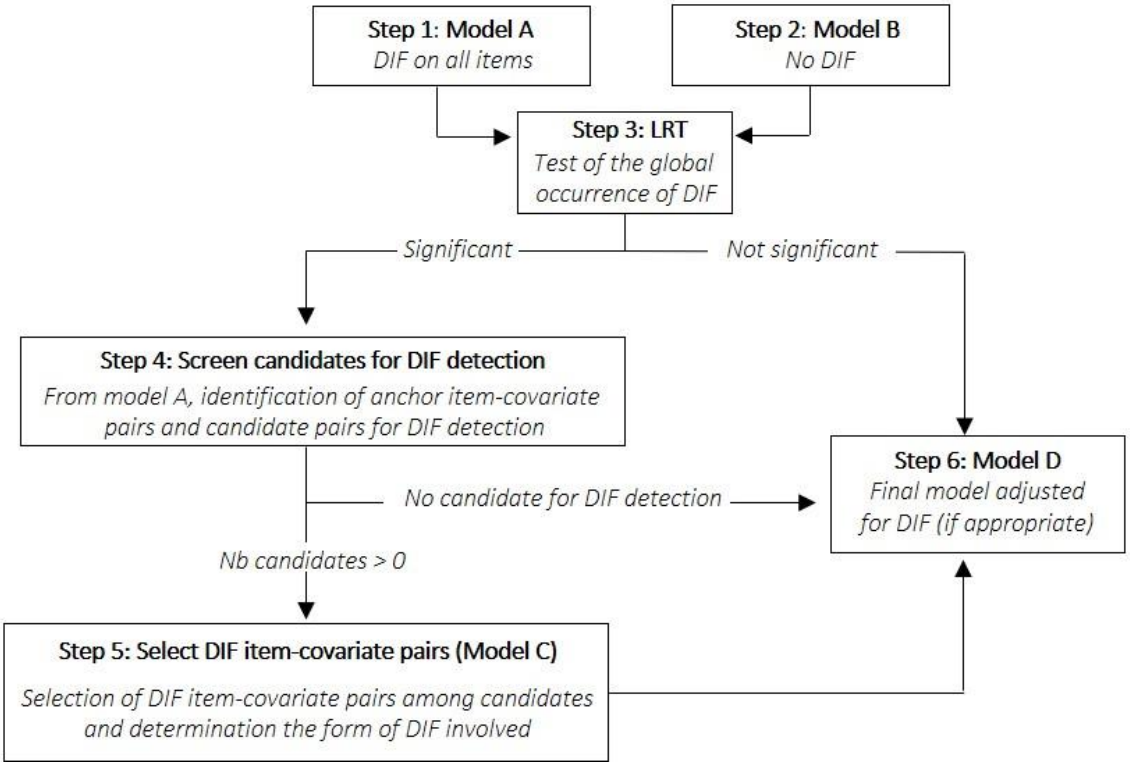
Notes: C designates here indistinctly covariates C_1 or C_2

This algorithm estimates several PCMs derived from Equation 2 with marginal maximum likelihood estimation. For all PCMs, the variances of the latent variable distribution across groups are assumed equal.

Fig 3: Graphical representation of the two ROSALI-DIF algorithms (ROSALI-DIF FORWARD and ROSALI-DIF BACKWARD).

Notes:

DIF: Differential item functioning, LRT: Likelihood-ratio test, Nb: Number



Likelihood penalization approach

A DIF detection method for polytomous items using likelihood penalization of a PCM or a generalized PCM (GPCM) has been comprehensively described by Schauberger and Mair in (21). By using likelihood penalization, the authors translated DIF detection into a parameter selection problem and aimed to determine which DIF parameters $\gamma_{jp}^{(C)}$ are worth estimating. The number of DIF parameters to be estimated depends entirely on the choice of the tuning parameter which controls the strength of the penalization. When this parameter is equal to zero, all DIF parameters are estimated (no penalization). On the contrary, no DIF parameters are estimated when this parameter tends to +∞.

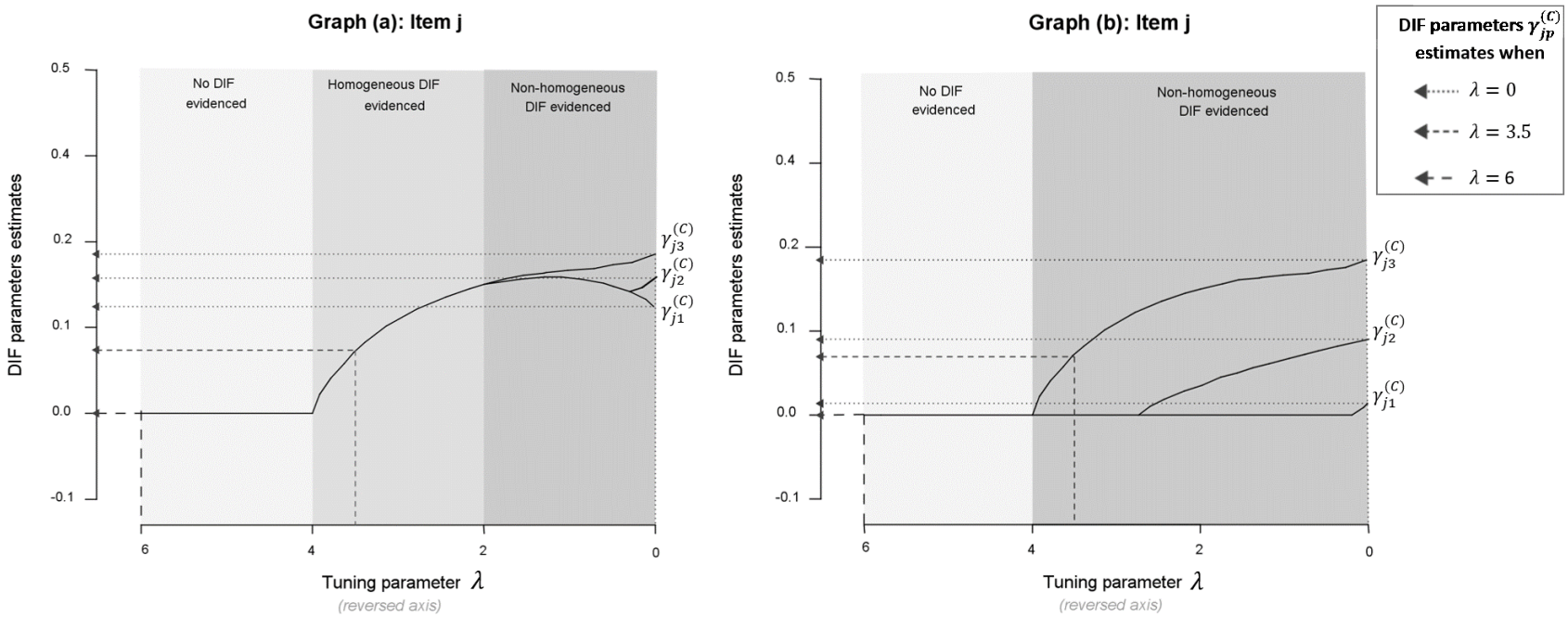
In practice, PCMs (or GPCMs) are estimated across a wide range of tuning parameters (one model for each value of tuning parameter). All estimates of the DIF parameters $\gamma_{jp}^{(C)}$ related to a given item-covariate pair are then plotted as functions the tuning parameter values in graphs called *DIF parameters paths*. Two main situations can arise when searching for DIF as pictured in Figure 4: either the graph shows an area where the DIF parameters $\gamma_{jp}^{(C)}$ are estimated but constrained to be equal (graph (a)), or the graph does not show such an area (graph (b)). The form of DIF evidenced is then entirely determined by the choice of the tuning parameter. The optimal tuning parameter is chosen to minimize the Bayesian information criterion.

We chose to evaluate this approach, named PCMLasso, enabling the detection of both forms of DIF using a PCM. Of note, the PCMLasso approach relies on a PCM where the item discrimination parameters are constrained to be equal over all items. In the ROSALI-DIF algorithms, the discrimination parameters equal 1 for all items. Besides, neither the ROSALI-DIF algorithms nor PCM-Lasso enable to consider that the DIF effect of one covariate may depend on the level of another covariate.

Fig 4:

Graph (a): DIF parameters $\gamma_{jp}^{(C)}$ paths with the PCMLasso approach for a given item-covariate pair in a fictious example (configuration A: the graph shows an area where the DIF parameters $\gamma_{jp}^{(C)}$ from Equation 2 are estimated but constrained to be equal). When the tuning parameter is large (left of the graph: $\lambda \in [6,4]$), no DIF parameter is estimated since the penalization is strong. In the middle of the graph ($\lambda \in]4,2]$), the penalization is not as strong, allowing the estimation of all DIF parameters (constrained to be equal over the response categories). Finally, at the right of the graph ($\lambda \in]2,0]$), the penalization is weak and the DIF parameters are no longer constrained to be equal over the response categories. If the optimal tuning parameter λ falls in the area $[6,4]$, then no DIF is evidenced for the item-covariate pair considered. If it falls in the area $]4,2]$ (respectively $]2,0]$), then homogeneous (respectively non-homogeneous) DIF is evidenced for the given pair.

Graph (b): DIF parameters paths with the PCMLasso approach for a given item-covariate pair (in this configuration B, the graph does not show an area where the DIF parameters are estimated but constrained to be equal, i.e., it does not show an area where homogeneous DIF could be evidenced). When the tuning parameter is large (left of the graph: $\lambda \in [6,4]$), no DIF parameter is estimated since the penalization is strong. Then, the penalization decreases, allowing the estimation of a first DIF parameter ($\gamma_{j3}^{(C)}$), then a second ($\gamma_{j2}^{(C)}$) and finally a third one ($\gamma_{j1}^{(C)}$). If the optimal tuning parameter λ falls in the area $[6,4]$, then no DIF is evidenced for the item-covariate pair considered, otherwise, non-homogeneous DIF is evidenced.



Simulation study

Data simulation

We simulated the responses of $n = 400$ or 800 individuals to a unidimensional questionnaire composed of $J = 4$ or 7 polytomous items (item 1, …, item J) with $M = 4$ response categories, numbered from 0 to $M - 1$. Individual latent variable levels were drawn from a standard normal distribution and responses were generated by a PCM. Item threshold parameters δ_{jp} were chosen to cover all the latent variable continuum (values are given in online appendix B). Items were numbered from 1 to J so that $\overline{\delta_{1p}} < \overline{\delta_{2p}} < \cdots < \overline{\delta_{Jp}}$ (where $\overline{\delta_{jp}}$ stands for the average value of the item j threshold parameters).

DIF operationalization

DIF was operationalized as between-group differences in item threshold parameters (groups being defined by observed covariates). For this simulation study, two binary covariates (denoted C_1 and C_2) were considered as possibly inducing DIF on items. Three settings were derived:

Setting No.1: The two covariates were not correlated and they each induced DIF on a different item.

Setting No.2: The two covariates were not correlated and they induced DIF on the same item.

Setting No.3: The two covariates were correlated and only one (i.e., C_1) induced DIF on two items. Further details on the process used to obtain correlated covariates are available in online appendix B alongside the formulas used for the DIF-items threshold parameters among each setting.

For each setting, two different forms of DIF were explored: homogeneous and non-homogeneous DIF. Homogeneous DIF was operationalized as between-group differences in item threshold parameters with the same direction and magnitude across the response categories (i.e., $\forall\, p,\ \gamma_{jp}^{(C)} = \gamma_j^{(C)}$, Figure 2.a). Non-homogeneous DIF was operationalized as between-group differences in item threshold parameters that varied in magnitude and/or direction. In our simulations, we only simulated between-group differences having the same direction but different magnitudes. Specifically, we shifted item threshold parameters by increasing values (e.g., Figure 2.b). Finally, we varied the DIF size (weak or medium). A comprehensive summary of the simulation study appears in Table 2. The combination of all simulation parameters led to 48 scenarios. We added 8 scenarios with no DIF as control scenarios. Each scenario was replicated 500 times and resulting datasets were then analyzed with the three DIF detection procedures (i.e., ROSALI-DIF algorithms and PCMLasso).

Table 2: Simulation plan summary

Questionnaire and sample features	
Number of items (J)	J = 4, 7 items
Number of response categories (M)	M = 4 response categories per item
Sample size (n)	n = 400, 800 simulated individuals
Latent variable (Θ)	
Mean μ, Variance σ^2	$\mu = 0, \sigma^2 = 1$
Main effect of covariates on latent variable level	No main effect of covariate C_1: $\beta_1 = 0$ No main effect of covariate C_2: $\beta_2 = 0$
DIF inducing covariates	
Setting No. 1	Covariates C_1 and C_2 are uncorrelated, they each induce DIF on a different item
Setting No. 2	Covariates C_1 and C_2 are uncorrelated, they both induce DIF on the same item
Setting No. 3	Covariates C_1 and C_2 are correlated, only covariate C_1 induces DIF on two items
DIF items	
J = 4 items	
Setting No. 1	Covariate C_1 induces DIF on item 2 and covariate C_2 induces DIF on item 3
Setting No. 2	Covariates C_1 and C_2 induce DIF on item 2
Setting No. 3	Covariate C_1 induces DIF on items 2 and 3
J = 7 items	
Setting No. 1	Covariate C_1 induces DIF on item 3 and covariate C_2 induces DIF on item 5
Setting No. 2	Covariates C_1 and C_2 induce DIF on item 3
Setting No. 3	Covariate C_1 induces DIF on items 3 and 5
DIF form	
Homogeneous	Same effect (magnitude and direction) of the covariate on all item threshold parameters: $\forall\, p,\ \gamma_{jp}^{(C)} = \gamma_j^{(C)}$
Non-homogeneous	The covariate has varying effects across the item threshold parameters
DIF size	
Weak	
Homogeneous DIF	$\forall\, p,\ \gamma_{jp}^{(C)} = \gamma_j^{(C)} = 0.3$
Non Homogeneous DIF	$\gamma_{j1}^{(C)} = 0.1, \gamma_{j2}^{(C)} = 0.3, \gamma_{j3}^{(C)} = 0.5$
Medium	
Homogeneous DIF	$\forall\, p,\ \gamma_{jp}^{(C)} = \gamma_j^{(C)} = 0.5$
Non Homogeneous DIF	$\gamma_{j1}^{(C)} = 0.1, \gamma_{j2}^{(C)} = 0.5, \gamma_{j3}^{(C)} = 0.9$

Notes: $C = C_1$ or C_2

Evaluation criteria

The performance of the three procedures in terms of DIF detection were evaluated according to different criteria.

Firstly, the rate of false detection of DIF among scenarios with no simulated DIF was computed as the proportion of datasets where DIF was wrongly detected on at least one item-covariate pair at the end of the procedures. We expected this rate to be low, but with no predefined threshold. As the

ROSALI-DIF algorithms involve a LRT performed at the 5% significance level, we also considered the proportion of datasets with a significant LRT to confront it to the nominal rate of 5%. A difference between the proportion of datasets with a significant LRT and the rate of false detection of DIF indicates that for some datasets, overall occurrence of DIF was initially suspected following the LRT, but finally not retained at the end of the procedure.

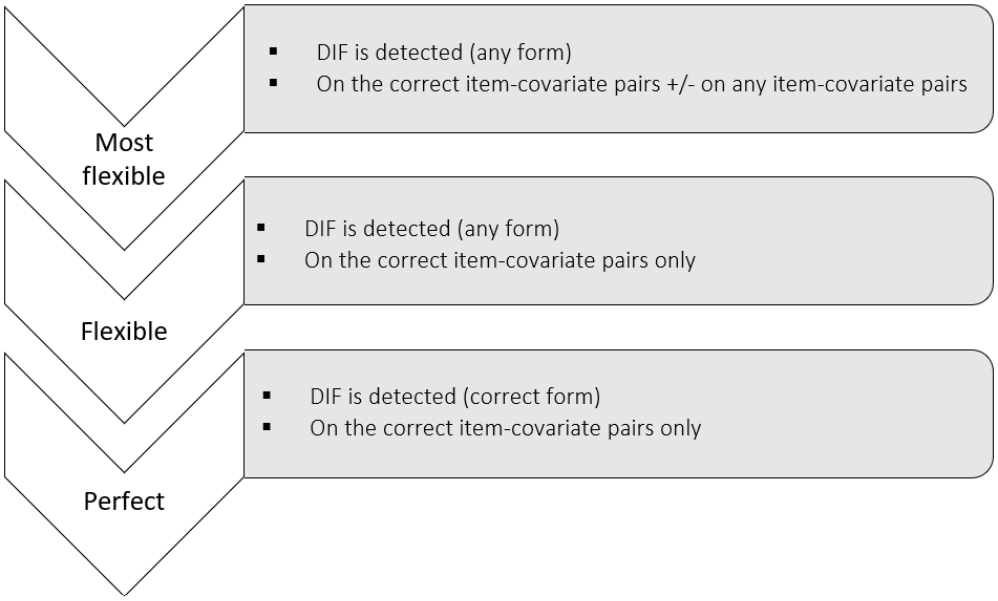
Secondly, among scenarios with simulated DIF, we used a set of criteria to assess the performance of the different procedures to detect DIF. They are given in Figure 5 by increasing level of requirement:

Most Flexible criterion: Did the procedure detect DIF on at least the correct item-covariate pairs (i.e., the pairs on which DIF was simulated)?

Flexible criterion: Did the procedure detect DIF on the correct item-covariate pairs only?

Perfect criterion: Did the procedure exactly detect what was simulated (i.e., DIF detected only on the correct item-covariate pair(s) and form of DIF rightly determined)?

Fig. 5: Criteria used to evaluate the DIF detection performance of the different procedures



The performance of the ROSALI-DIF algorithms and PCMLasso were assessed using the proportion of datasets meeting the Most Flexible, Flexible, and Perfect criteria at the end of the procedures. Although there is no predefined threshold, high proportions of datasets meeting the different criteria indicate good performance of the different procedures.

Subsequently, we studied the difference between the proportions of datasets meeting the different criteria. For a given procedure, the proportion of datasets meeting the Most flexible criterion but not meeting the Flexible criterion indicates in what proportion the procedure has identified more item-

covariate pairs affected by DIF than simulated. Similarly, the proportion of datasets meeting the Flexible criterion but not meeting the Perfect criterion indicates in which proportion the procedure detected the correct item-covariate pair(s) on which DIF was simulated (and only these) but failed to identify the form of DIF involved.

Finally, we assessed the bias in the estimation of the covariates' effects on the latent variable level (β_1 and β_2 in Eq.2) to determine whether the three methods enable for an unbiased estimation after DIF detection. In addition to bias, we computed the standard deviation of the $\beta_{k\ (k=1,2)}$ estimates. and the average model standard errors. We also compared the estimates of the DIF parameters $\widehat{\gamma_{Jp}^{(C)}}$ with the true simulated values using boxplots.

Stata software release 16 was used for data generation (*simirt* module, version 4.3 (36)). Analyses were performed using either Stata 16 for ROSALI-DIF algorithms or R 4.1.0 for the PCMLasso approach (*GPCMLasso* package version 0.1-5).

Results

Rates of false DIF detection – No DIF scenarios

Table 3 presents the rates of false DIF detection for both ROSALI-DIF FORWARD and PCMLasso. It additionally gives the proportion of datasets with a significant LRT for ROSALI-DIF FORWARD.

The proportions of datasets where DIF was wrongly detected on at least one item-covariate pair at the end of the algorithm were low for ROSALI-DIF FORWARD (from 3% to 6%). Neither the sample size n nor the number of items J seemed to impact these rates. However, rates of false detection of DIF were usually slightly lower when the two covariates C_1 and C_2 were correlated than when they weren't. For each scenario, the rate of false detection of DIF was systematically lower than the rate of datasets with a significant LRT. It means that for some datasets (1% to 3%), overall occurrence of DIF was initially suspected following the LRT, but finally not retained at the end of the algorithm. Of note, the rates of significant LRT were close to 5%, the nominal rate used for these tests. Same results were obtained for ROSALI-DIF BACKWARD (online appendix C).

DIF was wrongly detected on at least one item-covariate pair at the end of the PCMLasso in almost half of the datasets whatever the scenario. None of the simulation characteristics seemed to have an impact on these results. Despite the high rates of false DIF detection observed with PCMLasso, we chose to continue the investigations in order to draw a comprehensive view of its performance.

Table 3: Rates of false detection of DIF among scenarios with no simulated DIF computed at the end of each procedure (%DIF wrongly Detected) and rates of significant likelihood-ratio tests (%LRT SIG). Results are given according to the simulation characteristics n (sample size), *J* (number of items) and the presence or absence of correlation between covariates C_1 and C_2

			ROSALI-DIF FORWARD		PCMLasso
n	*J*	Corr	%LRT SIG	%DIF wrongly detected	%DIF wrongly detected
400	4	No	5%	4%	50%
400	4	Yes	6%	3%	50%
400	7	No	7%	6%	44%
400	7	Yes	6%	3%	54%
800	4	No	6%	4%	46%
800	4	Yes	5%	3%	47%
800	7	No	6%	5%	48%
800	7	Yes	4%	3%	46%

%DIF wrongly detected: Proportion of datasets where DIF was wrongly detected on at least one item-covariate pair at the end of the procedure (i.e., rate of false DIF detection)

%LRT SIG: Proportion of datasets with a significant likelihood-ratio test,

Corr: Correlation, indicates whether covariates C_1 and C_2 are correlated (=Yes) or not (=No)

Note: The procedures converged on all datasets. No identifiability issues were encountered.

Rates of correct DIF detection – DIF scenarios

Table 4 presents the proportion of datasets meeting the Most Flexible, Flexible and Perfect criteria at the end of the procedures for ROSALI-DIF FORWARD and PCMLasso. Results for ROSALI-DIF BACKWARD appear in online appendix C.

The performance of the three procedures varied depending on the values of the simulation characteristics. First, all procedures failed to detect DIF within scenarios with weak DIF and a sample size of 400: Most Flexible detection rates did not exceed 10% under these conditions. Therefore, results regarding these scenarios will not be further developed. The following paragraphs focus only on the results observed when the sample size n equals 800 or when the DIF size is medium (with either $n = 400$ or 800).

Table 4: Rates of correct DIF detection among DIF scenarios. Results are given according to the simulation characteristics: Setting, DIF form (Homogeneous H, Non-homogeneous NH), DIF size, sample size n, number of items J

					ROSALI-DIF FORWARD				PCMLasso		
Setting	DIF form	DIF size	n	J	%LRT SIG	Most flexible	Flexible	Perfect	Most flexible	Flexible	Perfect
1	H	Weak	400	4	31%	4%	3%	2%	5%	3%	0%
1	H	Weak	400	7	27%	5%	4%	3%	6%	3%	0%
1	H	Weak	800	4	64%	20%	17%	14%	16%	9%	1%
1	H	Weak	800	7	63%	26%	18%	14%	18%	9%	0%
1	H	Medium	400	4	85%	39%	34%	29%	35%	17%	0%
1	H	Medium	400	7	79%	44%	30%	25%	39%	16%	0%
1	H	Medium	800	4	99%	90%	73%	67%	80%	42%	1%
1	H	Medium	800	7	99%	91%	66%	59%	83%	46%	2%
1	NH	Weak	400	4	33%	5%	4%	1%	8%	6%	5%
1	NH	Weak	400	7	31%	4%	2%	0%	6%	3%	3%
1	NH	Weak	800	4	68%	29%	25%	3%	27%	18%	17%
1	NH	Weak	800	7	62%	27%	18%	2%	23%	14%	12%
1	NH	Medium	400	4	91%	56%	49%	6%	49%	26%	25%
1	NH	Medium	400	7	79%	51%	35%	5%	49%	24%	23%
1	NH	Medium	800	4	100%	96%	77%	31%	88%	49%	49%
1	NH	Medium	800	7	100%	96%	65%	24%	90%	52%	51%
2	H	Weak	400	4	29%	3%	1%	1%	5%	3%	0%
2	H	Weak	400	7	26%	2%	2%	1%	6%	2%	0%
2	H	Weak	800	4	68%	23%	20%	18%	17%	9%	0%
2	H	Weak	800	7	58%	21%	13%	10%	17%	8%	0%
2	H	Medium	400	4	83%	41%	35%	30%	35%	15%	0%
2	H	Medium	400	7	82%	41%	30%	25%	43%	15%	1%
2	H	Medium	800	4	99%	91%	77%	68%	76%	37%	1%
2	H	Medium	800	7	99%	92%	65%	61%	85%	46%	2%
2	NH	Weak	400	4	34%	6%	6%	1%	10%	6%	6%
2	NH	Weak	400	7	32%	7%	5%	1%	9%	5%	4%
2	NH	Weak	800	4	74%	34%	32%	2%	26%	16%	15%
2	NH	Weak	800	7	67%	32%	22%	2%	27%	15%	13%
2	NH	Medium	400	4	86%	50%	42%	5%	47%	27%	26%
2	NH	Medium	400	7	79%	51%	39%	6%	46%	22%	22%
2	NH	Medium	800	4	100%	95%	81%	26%	86%	54%	53%
2	NH	Medium	800	7	99%	96%	65%	22%	89%	51%	50%
3	H	Weak	400	4	21%	2%	2%	1%	2%	0%	0%
3	H	Weak	400	7	23%	2%	2%	2%	1%	1%	0%
3	H	Weak	800	4	44%	7%	7%	6%	2%	0%	0%
3	H	Weak	800	7	44%	13%	10%	8%	9%	3%	0%
3	H	Medium	400	4	58%	19%	18%	13%	7%	2%	0%
3	H	Medium	400	7	67%	26%	22%	18%	20%	6%	0%
3	H	Medium	800	4	91%	69%	61%	54%	19%	3%	0%
3	H	Medium	800	7	96%	78%	62%	56%	60%	21%	1%
3	NH	Weak	400	4	24%	3%	2%	0%	3%	1%	1%
3	NH	Weak	400	7	28%	2%	1%	0%	4%	1%	1%
3	NH	Weak	800	4	48%	13%	11%	1%	8%	4%	4%
3	NH	Weak	800	7	52%	17%	13%	2%	15%	6%	6%
3	NH	Medium	400	4	73%	31%	28%	5%	22%	8%	8%
3	NH	Medium	400	7	75%	38%	31%	6%	34%	16%	16%
3	NH	Medium	800	4	98%	86%	73%	25%	58%	19%	19%
3	NH	Medium	800	7	98%	85%	66%	26%	78%	34%	33%

%LRT SIG: Proportion of datasets with significant likelihood-ratio test, **Most flexible (%):** Proportion of datasets where the procedure identified DIF at least on the correct item-covariate pairs (among others), **Flexible (%):** Proportion of datasets where the procedure identified DIF on the correct item-covariate pairs only, **Perfect (%):** Proportion of datasets were the procedure identified exactly the DIF that was simulated (correct form and correct pairs)

Setting No. 1: The two covariates are not correlated and they induce DIF on two distinct items
Setting No. 2: The two covariates are not correlated and they induce DIF on the same item
Setting No. 3: The two covariates are correlated and only one induces DIF on two items
Note: The procedures converged on all datasets. No identifiability issues were encountered.

Performance within settings No. 1 and 2 (uncorrelated covariates both inducing DIF)

Whatever the procedure, Most Flexible detection rates were low when DIF size was weak and $n = 800$ (from 16% to 34%). However, when DIF size was medium, Most Flexible detection rates were moderate when $n = 400$ (between 35% and 56%) and high when $n = 800$ (from 76% to 96%). As a reminder, these rates indicate to what extent the different procedures are able to detect DIF on at least the item-covariate pairs on which DIF was simulated. Best performance regarding Most Flexible detection rates was observed for ROSALI-DIF FORWARD (ranging from 20% to 96%, mean: 56%) but ROSALI-DIF BACKWARD also showed quite similar performance (rates did not differ by more than 5%). The performances of PCMLasso were generally slightly lower (rates ranging from 16% to 90%, mean: 50%). Of note, the three methods showed usually higher Most Flexible detection rates when DIF was non-homogeneous than when it was homogeneous, all other scenario characteristics being equal, with a maximal difference up to +17% (mean difference of +8%).

Both ROSALI-DIF algorithms showed poor Flexible detection rates when DIF size was weak and $n = 800$ (from 13% to 32%), moderate Flexible detection rates when DIF size was medium and $n = 400$ (from 30% to 49%), and high Flexible detection rates when DIF size was medium and $n = 800$ (from 65% to 81% and 78% to 87% for ROSALI-DIF FORWARD and ROSALI-DIF BACKWARD, respectively). Regarding PCMLasso, we observed poor Flexible detection rates in all scenarios except those with a sample size of 800 and medium DIF size where rates ranged from 37% to 54%. Hence, based on the Flexible criteria, the best-performing methods are the ROSALI-DIF algorithms. Of note, ROSALI-DIF BACKWARD outperformed ROSALI-DIF FORWARD when DIF size was medium and $n = 800$ (Flexible detection rates of both methods differed from +5% to +16%) while their performance was similar under other scenarios. Finally, the three methods showed generally higher Flexible detection rates when DIF was non-homogeneous than when it was homogeneous (all other scenario characteristics being equal), with a maximal difference up to +17% (mean difference of +7%).

Flexible detection rates were lower than the Most Flexible detection rates whatever the procedure. It means that, in addition to the correct DIF item-covariate pair, all procedures wrongly detected other pairs (on which DIF was not simulated). Gaps between the Most Flexible and Flexible detection rates usually increased with increasing Most Flexible detection rates for all procedures (the higher the Most Flexible detection rate, the greater the gap). For both ROSALI-DIF algorithms, gaps also increased with increasing number of item J. These gaps were always the smallest for ROSALI-DIF BACKWARD and the largest for PCMLasso.

Among scenarios with homogeneous DIF, ROSALI-DIF algorithms both showed Perfect detection rates close to the Flexible detection rates (e.g., they differed from 4% to 10% for medium DIF scenarios). Thus, within datasets meeting the Flexible criteria, both algorithms correctly determined the form of

DIF involved when the simulated DIF was homogeneous. As for PCMLasso, Perfect detection rates did not exceed 2% indicating that PCMLasso failed to identify the correct DIF form when the simulated DIF was homogeneous. Hence, based on the Perfect criterion, both ROSALI-DIF algorithms outperformed PCMLasso within scenarios with homogeneous DIF. However, when the simulated DIF was non-homogeneous, Perfect detection rates associated with ROSALI-DIF algorithms were substantially lower than Flexible detection rates (e.g., gaps ranged from 30% to 56% among scenarios with medium DIF). Hence, both algorithms struggled to identify the correct form of DIF (i.e. non-homogeneous) among scenarios meeting the Flexible criteria. On the contrary, PCMLasso showed Perfect detection rates very close to the Flexible detection rates (i.e. rates ranging from 49% to 53% when DIF was medium and $n = 800$, low rates ranging from 12% to 26% otherwise). It indicates that once PCMLasso correctly identified the item-covariate pairs affected by DIF, it also correctly identified the DIF form. Therefore, when DIF was non-homogeneous, the PCMLasso approach showed larger Perfect detection rates than the ROSALI-DIF algorithms. Based on the Perfect criterion, PCMLasso outperformed ROSALI-DIF algorithms when the simulated DIF was non-homogeneous. Note that, as observed for Flexible detection rates, ROSALI-DIF BACKWARD showed higher Perfect detection rates than ROSALI-DIF FORWARD in scenarios with medium DIF and $n = 800$. Both algorithms showed similar performance otherwise.

Performance within setting No. 3 (correlated covariates with only one inducing DIF)

Among scenarios of setting No. 3, the performance of the three DIF detection methods were usually poorer as compared to settings No. 1 and 2 for all criteria.

Indeed, regarding the Most flexible and Flexible detection rates, the performance of all three methods was globally poor when DIF size was weak or when the sample size equaled 400. For both ROSALI-DIF algorithms, these rates increased among scenarios with n =800 and medium DIF (ranging from 69% to 86% for the Most Flexible detection rates and from 61% to 80% for the Flexible detection rate). Performance was poorer for PCMLasso under the same conditions as Most Flexible and Flexible detection rates ranged from 19% to 78% and from 3% to 34%, respectively. Of note, we observed similar effects to those highlighted in settings No. 1 and 2 regarding these rates and the associated gaps.

Bias and empirical standard error – DIF scenarios

Bias, empirical standard errors, and average model standard errors associated with the estimation of β_1 and β_2 (the respective effect of the covariates C_1 and C_2 on the latent variable level) are given in online appendix D. Under settings No. 1 and 2, bias remained small for all methods; it never exceeded 0.08 in absolute value. Under setting No. 3 results were more mixed: bias related to the estimation of β_1 (the effect of C_1, the only DIF-inducing covariate in this setting) remained small when $J = 7$ but it increased when $J = 4$ for all procedures (reaching -0.19 for ROSALI-DIF algorithms and -0.14 for PCMLasso). Across all settings, the better the methods detected DIF, the lower the bias.

DIF parameter estimates

On the one hand, PCMLasso always underestimated the DIF parameters $\gamma_{jp}^{(C)}$. On the other hand, DIF parameters estimates were much closer to the true simulated values for ROSALI-DIF algorithms but they showed a larger dispersion (see the boxplots available through the R Shiny app, online appendix E).

Discussion

Main results

This study aimed to extend the first part of the ROSALI algorithm dedicated to DIF detection using RMT to consider two binary covariates instead of one. We proposed two extensions: ROSALI-DIF FORWARD and ROSALI-DIF BACKWARD. The novelty characterizing these extensions is the screening step that aims to identify the item-covariate pairs candidate for DIF detection and the item-covariate pairs that will be considered as anchors (not affected by DIF). This further step was inspired by the iterative Wald test procedure proposed by Tay et al. (34,35). These authors indicated that testing all items for DIF in a fully unconstrained model had good power but a high Type I error, so it could be useful for identifying anchor items in a preliminary stage [34,35]. Performance of each extensions of ROSALI for DIF detection were assessed by simulations alongside the performance of the approach based on likelihood penalization proposed by Schauberger and Mair (21) (PCMLasso) under conditions that were aimed to be representative of health research contexts .

In light of the rates of false detection of DIF, both ROSALI-DIF algorithms satisfactorily prevent from inferring DIF when it has not been simulated. This good performance may be explained by: (i) combining the LRT performed at a 5% significance level with the screening and iterative steps, and (ii) the Bonferroni correction applied during the iterative step. In light of the Flexible and Most Flexible detection rates, both ROSALI-DIF algorithms can detect item-covariates pairs having medium DIF (as simulated in this manuscript) with good power for studies with two correlated or uncorrelated binary

covariates, a sample size of 800, and a questionnaire similar to the ones simulated with regards to M and J. Moreover, ROSALI-DIF algorithms should generally not wrongly detect items-covariates pairs without DIF. However, one must be cautious regarding the form of DIF evidenced by these algorithms, as non-homogenous DIF is rarely identified as such. It means that the test performed during step 5 generally lacked power as it failed to reject the null hypothesis of homogeneous DIF when DIF was actually non-homogenous. Correct identification of the DIF form may require larger sample sizes.

Regarding PCMLasso, rates of false detection of DIF among scenarios without DIF were high. Indeed, PCMLasso was prone to erroneously detect DIF in at least one item-covariate pair in almost half of the datasets, no matter the scenario. This drawback is also highlighted by the large gaps between the Flexible and Most flexible detection rates among scenarios where DIF was simulated. Indeed, PCMLasso was likely to wrongly flag other item-covariate pairs in addition to the ones truly affected by DIF. We noticed that the estimated size of the wrongly evidenced DIF remained small on average which may not result in a meaningful measurement bias at the scale level (data not shown). One must be cautious regarding the form of DIF evidenced by this approach. Indeed, it almost always suspected the occurrence of non-homogeneous DIF (whatever the form of DIF simulated). After further investigations, we also noticed that in some datasets, DIF parameters estimates were very close (e.g., $\gamma_{j1}^{(C)} = 0.28$ and $\gamma_{j2}^{(C)} = \gamma_{j3}^{(C)} = 0.29$), indicating that the tuning parameter may have been chosen just after the split in the DIF parameters path (see the border between the "Homogeneous DIF" area and the "Non-homogeneous DIF" area pictured in Figure 4.a).

All three methods showed small power to detect weak DIF. It may not be a major issue as the weak DIF simulated in this simulation study might generally not result in a meaningful measurement bias at the scale level. Moreover, DIF detection performance of ROSALI-DIF algorithms and PCMLasso decreased with decreasing sample size. This effect was expected but is sharply marked and may be even more problematic with unbalanced covariates.

The DIF parameters $\gamma_{jp}^{(C)}$, were satisfactorily recovered for ROSALI-DIF algorithms among datasets meeting the Most Flexible criteria. Conversely, the PCMLasso approach always underestimated them due to the penalization that shrinks them toward zero. This downward bias has already been previously highlighted [21,37]. Of note, Tutz and Shauberger indicated that the bias introduced by the parameter shrinkage could be removed by an additional refit (i.e., fitting a final unpenalized model that only includes DIF effects evidenced after the likelihood penalization approach).

Finally, all procedures provided a globally unbiased estimation of the effect of the covariates on the latent variable level adjusted for DIF. The most biased estimates were obtained for covariate C_1 under the setting No. 3 with a test composed of four items and half of them affected by DIF induced by a single

covariate. As demonstrated by Rouquette et al. in a simulation study, such a configuration may lead to a meaningful bias if DIF is ignored (6). Hence, the fact that the three procedures showed generally lower DIF detection performance among these scenarios may be one of the causes of these biased estimates. Of note, DeMars et al. reported that DIF is conceptualized as differences in the item endorsement probabilities after controlling the psychological variable or capacity targeted by the questionnaire (38). Yet, if a large proportion of items is affected by DIF (i.e. $\geq 50\%$), then the questionnaire might measure different constructs among the groups being compared, and it would make no sense to speak about controlling the psychological variable level or ability (38) as the target construct may not be conceptualized in the same way across groups.

Limitations and perspectives

Several limitations can be addressed. First, neither ROSALI extensions nor the PCM Lasso approach allows to consider that the DIF effect of one covariate may depend on the level of another covariate. To allow for such a phenomenon, it would be necessary to introduce an interaction term between the covariates. Yet, more developments are needed to know how to deal with such an interaction. Of note, the PCM-IFT approach systematically models such an interaction because of its philosophy of recursive partitioning. More broadly, it would be worth comparing our results with the performance of the PCM-IFT approach under the same conditions when it will be able to account for the effect of a covariate on the latent variable.

Besides, we have considered simple situations with two well-balanced covariates that did not impact the latent variable level. Extending the simulation framework with: i) less balanced covariates, ii) covariates with an impact on the latent variable level, and iii) more than two covariates, would be interesting to have more insight into the performance of the evaluated procedures. In addition, it would be worth evaluating these procedures under a wider DIF setting (e.g., considering different DIF directions within one scenario). Finally, all datasets were complete: there were no missing data with regards to item responses nor covariate. A simulation study with incomplete datasets could be of value to evaluate the performance of the DIF detection methods, in a situation more representative of real data. Of note, as parameters are estimated using marginal maximum likelihood, estimations should be asymptotically unbiased in case of incomplete data missing (completely) at random. Nonetheless, due to a loss of precision of the estimations, a loss in the performance can be expected.

The methods evaluated in this manuscript are based on statistical results. Hence, it is critical to ensure that evidenced DIF effects are relevant and meaningful. Besides, it could be interesting to add *a priori* clinical knowledge on the items on which DIF may be likely, depending on the patients' population in order to not only rely on statistical criteria. Further developments are needed to better grasp the

sources and determinants of lack of measurement invariance in health research. To that end, DIF detection methods allowing considering several covariates simultaneously (with two categories or more) could be of great practical interest.

Disclosure

Data availability statement: Simulated datasets, modules and scripts related to this simulation study are openly available at https://osf.io/hkp38/?view_only=81c3c3c42e194760a7accbd7cc9b431f

Declaration of conflicting interest: The Authors declare that there is no conflict of interest.

Funding: YD received a national grant from the French Ministry of Higher Education, Research and Innovation.

References

1. Basch E. Patient-Reported Outcomes — Harnessing Patients' Voices to Improve Clinical Care. N Engl J Med. 2017 Jan 12;376(2):105–8.
2. Sawatzky R, Chan EKH, Zumbo BD, Ahmed S, Bartlett SJ, Bingham CO, et al. Montreal Accord on Patient-Reported Outcomes (PROs) use series-Paper 7: modern perspectives of measurement validation emphasize justification of inferences based on patient reported outcome scores. J Clin Epidemiol. 2017 Sep;89:154–9.
3. Millsap RE. Statistical approaches to measurement invariance. New York: Routledge; 2011. 355 p.
4. Mellenbergh GJ. Item bias and item response theory. Int J Educ Res. 1989 Jan;13(2):127–43.
5. Millsap RE, Everson HT. Methodology Review: Statistical Approaches for Assessing Measurement Bias. Appl Psychol Meas. 1993 Dec;17(4):297–334.
6. Rouquette A, Hardouin JB, Coste J. Differential Item Functioning (DIF) and Subsequent Bias in Group Comparisons using a Composite Measurement Scale: A Simulation Study. J Appl Meas. 2016;17(3):312–34.
7. Vanier A, Oort FJ, McClimans L, Ow N, Gulek BG, Böhnke JR, et al. Response shift in patient-reported outcomes: definition, theory, and a revised model. Qual Life Res Int J Qual Life Asp Treat Care Rehabil. 2021 Dec;30(12):3309–22.
8. Sprangers MAG, Schwartz CE. Integrating response shift into health-related quality of life research: a theoretical model. Soc Sci Med. 1999 Jun;48(11):1507–15.
9. Holland PW, Thayer DT. Differential item performance and the Mantel-Haenszel procedure. In: Test validity. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc; 1988. p. 129–45.
10. Rogers HJ, Swaminathan H. A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Appl Psychol Meas. 1993;17(2):105–16.

11. Thissen D, Steinberg L, Gerrard M. Beyond group-mean differences: The concept of item bias. Psychol Bull. 1986;99(1):118–28.

12. Thissen D, Steinberg L, Wainer H. Use of item response theory in the study of group differences in trace lines. In: Test validity. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc; 1988. p. 147–72.

13. Thissen D, Steinberg L, Wainer H. Detection of differential item functioning using the parameters of item response models. In: Differential item functioning. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc; 1993. p. 67–113.

14. Lord FM. Applications of item response theory to practical testing problems. Repr. New York: Routledge; 2008. 274 p.

15. Jones RN. Differential Item Functioning and its Relevance to Epidemiology. Curr Epidemiol Rep. 2019 Jun;6(2):174–83.

16. Tay L, Vermunt JK, Wang C. Assessing the Item Response Theory With Covariate (IRT-C) Procedure for Ascertaining Differential Item Functioning. Int J Test. 2013 Jul;13(3):201–22.

17. Tay L, Huang Q, Vermunt JK. Item Response Theory With Covariates (IRT-C): Assessing Item Recovery and Differential Item Functioning for the Three-Parameter Logistic Model. Educ Psychol Meas. 2016 Feb;76(1):22–42.

18. El-Komboz BA, Zeileis A, Strobl C. Detecting Differential Item and Step Functioning with Rating Scale and Partial Credit Trees. :34.

19. Bollmann S, Berger M, Tutz G. Item-Focused Trees for the Detection of Differential Item Functioning in Partial Credit Models. Educ Psychol Meas. 2018 Oct;78(5):781–804.

20. Oberski DL, van Kollenburg GH, Vermunt JK. A Monte Carlo evaluation of three methods to detect local dependence in binary data latent class models. Adv Data Anal Classif. 2013 Sep;7(3):267–79.

21. Schauberger G, Mair P. A regularization approach for the detection of differential item functioning in generalized partial credit models. Behav Res Methods. 2020 Feb;52(1):279–94.

22. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 1992 Jun;30(6):473–83.

23. Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatr Scand. 1983 Jun;67(6):361–70.

24. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. J Clin Epidemiol. 2010 Nov 1;63(11):1179–94.

25. Hammas K, Sébille V, Brisson P, Hardouin JB, Blanchin M. How to Investigate the Effects of Groups on Changes in Longitudinal Patient-Reported Outcomes and Response Shift Using Rasch Models. Front Psychol. 2020;11:3704.

26. Blanchin M, Guilleux A, Hardouin JB, Sébille V. Comparison of structural equation modelling, item response theory and Rasch measurement theory-based methods for response shift detection at item level: A simulation study. Stat Methods Med Res. 2020 Apr;29(4):1015–29.

27. Blanchin M, Brisson P, Sébille V. Performance of a Rasch-based method for group comparisons of longitudinal change and response shift at the item level in PRO data: A simulation study. Methods. 2022 Aug;204:327–39.

28. Fischer GH, Molenaar IW, editors. Rasch Models [Internet]. New York, NY: Springer New York; 1995 [cited 2021 Nov 10]. Available from: http://link.springer.com/10.1007/978-1-4612-4230-7

29. Masters GN. A rasch model for partial credit scoring. Psychometrika. 1982 Jun;47(2):149–74.

30. Fischer GH, Ponocny I. An extension of the partial credit model with an application to the measurement of change. Psychometrika. 1994 Jun;59(2):177–92.

31. Christensen KB, Kreiner S, Mesbah M. Rasch Models in Health. John Wiley & Sons; 2012. 24 p.

32. Penfield RD. Assessing Differential Step Functioning in Polytomous Items Using a Common Odds Ratio Estimator. J Educ Meas. 2007;44(3):187–210.

33. Penfield R, Alvarez K, Lee O. Using a Taxonomy of Differential Step Functioning to Improve the Interpretation of DIF in Polytomous Items: An Illustration. Appl Meas Educ. 2009 Jan 1;22:61–78.

34. Tay L, Meade AW, Cao M. An Overview and Practical Guide to IRT Measurement Equivalence Analysis. Organ Res Methods. 2015 Jan 1;18(1):3–46.

35. Cao M, Tay L, Liu Y. A Monte Carlo Study of an Iterative Wald Test Procedure for DIF Analysis. Educ Psychol Meas. 2017 Jan;77(1):104–18.

36. Hardouin JB. SIMIRT: Stata module to process data generated by IRT models [Internet]. Statistical Software Components. Boston College Department of Economics; 2013 [cited 2021 Dec 15]. Available from: https://ideas.repec.org/c/boc/bocode/s450402.html

37. Tutz G, Schauberger G. A Penalty Approach to Differential Item Functioning in Rasch Models. Psychometrika. 2015 Mar;80(1):21–43.

38. DeMars CE, Lau A. Differential Item Functioning Detection With Latent Classes: How Accurately Can We Detect Who Is Responding Differentially? Educ Psychol Meas. 2011 Aug;71(4):597–616.

DOCTORAT BIOLOGIE BRETAGNE SANTE LOIRE

Nantes Université

Titre : Développement et évaluation de méthodes permettant d'étudier les sources d'hétérogénéité du DIF et du Response Shift lors de l'analyse des données rapportées par les patients

Mots clés : Données rapportées par les patients, Invariance de la mesure, Response Shift, DIF, Psychométrie

Résumé : Les concepts subjectifs comme la fatigue, l'anxiété ou encore la qualité de vie font l'objet d'un intérêt croissant en recherche biomédicale. Ces concepts subjectifs n'étant pas directement observables, on les mesure à l'aide de questionnaires généralement autorapportés par les patients. L'analyse de ce type de données peut néanmoins s'avérer complexe. En effet, la comparaison directe des scores obtenus à un questionnaire entre différents groupes d'individus ou au cours du temps suppose que tous les individus interprètent les items du questionnaire de la même façon et que cette interprétation ne change pas au cours du temps. Cette hypothèse peut s'avérer fausse. Par exemple, certains individus peuvent interpréter différemment le concept mesuré (et les questions associées) en raison de leurs caractéristiques culturelles, environnementales et personnelles, mais également à cause des expériences qu'ils ont vécues. On parle alors de fonctionnement différentiel des items (DIF). De plus, suite à un événement de santé important, certains individus peuvent également interpréter différemment les questions qui leur sont posées au cours du temps (response shift, RS). Le DIF et le RS doivent être pris en compte, car ils peuvent potentiellement fausser l'interprétation des données recueillies, mais également être des phénomènes d'intérêt en soi. L'objectif de cette thèse est de proposer et d'évaluer par simulations des développements méthodologiques pour étudier les sources d'hétérogénéité de ces deux phénomènes. Ce manuscrit est composé de trois travaux principaux : (i) Le développement d'une méthode visant à individualiser la détection du RS, (ii) le développement d'une méthode de détection du DIF en prenant simultanément en compte deux covariables binaires et (iii) l'étude des propriétés psychométriques de l'inventaire du développement post-traumatique (un phénomène potentiellement lié au RS).

Title: Development and evaluation of methods to examine sources of heterogeneity in DIF and response shift when analyzing patient-reported outcome data

Keywords: Patient-reported outcomes, Measurement invariance, Response Shift, DIF; Psychometrics

Abstract: Subjective concepts such as fatigue, anxiety, or quality of life are increasingly being used in biomedical research. These unobservable subjective concepts are measured by means of questionnaires, usually self-reported by patients. However, the analysis of these data can be challenging. Indeed, the direct comparison of questionnaire scores between different groups of individuals or over time assumes that all individuals interpret the questions in the same way and that this interpretation does not change over time. Yet, this assumption may be unmet. For example, some individuals may interpret the measured concept (and the associated items) differently because of their cultural, environmental, and personal characteristics, but also because of the experiences they went through. This phenomenon is known as differential item functioning (DIF). In addition, following a major health event, some individuals may also change their interpretation of the items composing the questionnaire (response shift, RS). Both DIF and RS can threaten the accurate interpretation of patient-reported outcome measures. In addition, they are also phenomena that deserve to be studied. This thesis aims to propose and evaluate, by simulations, methodological developments to study the sources of heterogeneity of these two phenomena. This manuscript is composed of three main works: (i) the development of a method aiming at individualizing RS detection, (ii) the development of a method for detecting DIF in the presence of two binary covariates, and (iii) the study of the psychometric properties of the post-traumatic growth inventory (a phenomenon potentially related to RS).