Statistical inference for categorization models and presentation order

Giulia Mezzadri

UNIVERSITÉ CÔTE D'AZUR
École Doctorale de Sciences Fondamentales et Appliquées

THÈSE DE DOCTORAT

En vue de l'obtention du grade de

Docteur en Mathématiques

de l'UNIVERSITÉ CÔTE D'AZUR

Présentée et soutenue par

Giulia MEZZADRI

INFÉRENCE STATISTIQUE POUR LES MODÈLES DE CATÉGORISATION ET ORDRE DE PRÉSENTATION

STATISTICAL INFERENCE FOR CATEGORIZATION MODELS AND PRESENTATION ORDER

Thèse dirigée par **Patricia REYNAUD-BOURET**, **Fabien MATHY** et **Thomas LALOË**

Soutenue le 11 Décembre 2020

Au vu des rapports de

Royce ANDERS	Maître de conférences HDR	Université de Lyon 2	*Rapporteur*
Susanne DITLEVSEN	Professeur	University of Copenhagen	*Rapportrice*

Jury

Royce ANDERS	Maître de conférences HDR	Université de Lyon 2	*Rapporteur*
Robert GOLDSTONE	Professeur	Indiana University	*Examinateur*
Thomas LALOË	Maître de conférences HDR	Université Côte d'Azur	*Co-encadrant de thèse*
Fabien MATHY	Professeur	Université Côte d'Azur	*Co-directeur de thèse*
Pierre PUDLO	Professeur	Université d'Aix-Marseille	*Examinateur*
Olivier RENAUD	Professeur	Université de Genève	*Examinateur*
Patricia REYNAUD-BOURET	Directeur de Recherche	CNRS	*Directrice de thèse*

INFÉRENCE STATISTIQUE POUR LES MODÈLES DE CATÉGORISATION ET ORDRE DE PRÉSENTATION

Résumé

Cette thèse se consacre à l'étude de la catégorisation, qui est la capacité cognitive de placer des objets dans des groupes. Plus particulièrement, nous nous intéressons à l'ordre de présentation et à la modélisation. L'objectif de cette thèse est triple : étudier l'influence de l'ordre de présentation dans l'apprentissage des catégories ; fournir une méthode statistique robuste pour la comparaison des modèles de catégorisation ; et déterminer si les modèles de catégorisation sont sensibles à différents types d'ordre.

Dans un premier temps, nous décrivons les expériences effectuées dont le but est d'explorer l'influence de deux types d'ordre sur la vitesse d'apprentissage. Dans ces expériences, les participants devaient apprendre une règle de catégorisation concernant des objets quadridimensionnels. Les stimuli étaient présenté en utilisant soit un ordre par règle, qui présente les objets de la règle principale en premier, soit un ordre par similarité, qui maximise la similarité entre objets consécutifs. Nous trouvons que l'ordre par règle facilite l'apprentissage quand les stimuli sont présentés dans le même ordre d'un bloc à un autre et quand les catégories sont bloquées ou présentées de façon aléatoire.

Dans un deuxième temps, nous décrivons les modèles de catégorisation utilisés dans cette thèse et introduisons un nouveau modèle de catégorisation qui intègre l'ordre de présentation. Parmi les modèle de transfert, qui ne sont adaptés qu'à reproduire la phase de transfert, nous décrivons le Generalized Context Model (GCM), qui est un des modèles de catégorisation les plus connus, le Generalized Context Model équipé avec un mécanisme de délai (GCM-Lag), qui est une extension du GCM qui prend en compte un mécanisme de perte de mémoire, et le nouveau Ordinal General Context Model (OGCM), qui est une extension du GCM qui prend en compte l'ordre de présentation des stimuli. Parmi les modèles d'apprentissage, qui reproduisent aussi bien la phase d'apprentissage que celle de transfert, nous décrivons ALCOVE, qui intègre la logique du GCM dans un réseau de neurones, et Component-Cue, qui intègre une stratégie par induction dans un réseau de neurones.

Dans un troisième temps, nous développons une méthode robuste d'inférence statistique pour comparer les modèles de catégorisation. Cette méthode est séparément appliquée aux modèles de transfert et d'apprentissage. Nous trouvons que le modèle de transfert qui reproduit au mieux notre jeu de données est la version d'OGCM qui prend en

considération l'ordre de présentation le plus fréquent pendant la phase d'apprentissage. Ce résultat montre que l'information fournie par la composante temporelle est importante pour classifier les objets. De plus, nous trouvons que Component-Cue capture mieux les motifs de généralisation et ALCOVE capture mieux la dynamique d'apprentissage. Enfin, nous explorons la possibilité que les performances des modèles d'apprentissage soient dépendantes de l'ordre de présentation. Nous montrons que ALCOVE et Component-Cue sont sensibles à l'ordre et que les motifs de généralisation de Component-Cue sont compatibles avec un apprentissage par règle.

Pour terminer, nous décrivons comment appliquer des modèles de transfert à des données d'apprentissage en utilisant la segmentation et la segmentation/clustering. L'application de ces deux techniques à notre jeu de données montre la présence de deux groupes d'individus : les individus rapides et les individus lents. De plus, en utilisant la segmentation/clustering, nous trouvons une relations entre la vitesse d'apprentissage des individus et l'ordre de présentation. In particulier, les individus qui ont montré une vitesse d'apprentissage élevée au début de l'expérience avait reçu pour la plupart un ordre par règle.

Mots clés : catégorisation, ordre de présentation, ordre par règle, ordre par similarité, modèles de catégorisation, ALCOVE, Component-Cue, GCM, inférence statistique, segmentation, segmentation/clustering.

STATISTICAL INFERENCE FOR CATEGORIZATION MODELS AND PRESENTATION ORDER

Abstract

This thesis is devoted to the study of categorization, which is the cognitive ability to organize entities into groups. In particular, we focus on presentation order and modeling. The objective is threefold: to investigate the influence of presentation order on category learning; to provide a robust statistical method to compare categorization models; and to explore whether categorization models are sensitive to specific types of order.

Firstly, we report several experiments exploring the effects of two types of order on learning speed. In our experiments, participants were taught 4-feature category structures using either a rule-based presentation order (in which exemplars are presented following a "principal rule plus exceptions" structure) or a similarity-based presentation order (which maximizes the similarity between contiguous exemplars). We find that the rule-based order facilitates learning when the across-blocks manipulation was constant and categories were either blocked or randomly alternated.

Secondly, we describe the selected categorization models and introduce a new categorization model that integrates the order in which stimuli are presented. Among transfer models (which are adapted to reproduce participants' performance during the transfer phase), we describe the Generalized Context Model (GCM), one of the most well-known categorization models; the Generalized Context Model equipped with a Lag mechanism (GCM-Lag), an extension of the GCM integrating a memory decay mechanism; and the new Ordinal General Context Model (OGCM), an extension of the GCM incorporating temporal information. Among learning models (which can reproduce participants' performance during both the learning and transfer phases), we describe ALCOVE, which integrates the logic underlying the GCM in a neural network structure, and Component-Cue, which incorporates an induction strategy in a neural network structure.

Thirdly, we develop a robust statistical method for comparing categorization models and apply it separately to both transfer and learning models. We find that the transfer model that best fits our dataset is the version of the OGCM that takes into account the most frequent presentation order received during the learning phase. This result indicates that the information provided by the ordinal dimension is relevant for the categorization task. We also find that Component-Cue best captures the generalization patterns, while ALCOVE best captures the learning dynamics. Finally, we investigate whether the performance of

the selected learning models is related to the type of order in which stimuli are presented. We find that both learning models are sensitive to presentation order, in particular the generalization patterns of Component-Cue are consistent with a rule-based retrieval.

Finally, we describe how to apply transfer models to learning data using segmentation and segmentation/clustering. The application of these two techniques to our dataset shows that there are two groups of participants: high-speed and low-speed. Moreover, using the segmentation/clustering method, we find a relation between participants' learning speed and type of order. In particular, participants that showed a high-speed learning in the early stage of the categorization task mostly received a rule-based order.

Keywords: categorization, presentation order, rule-based order, similarity-based order, categorization models, ALCOVE, Component-Cue, GCM, statistical inference, segmentation, segmentation/clustering.

Acknowledgement

What a crazy roller coaster has been this thesis! The health issues, the broken bones, the lack of a well-defined project, the willing to quit, the pandemic, but also the deep friendships, the delightful teaching experience, the gain of a scientific autonomy, the decision to fight, the eclectic learning, the discovery of my non-binary identity, the personal and professional growth. If I could turn back time, I would not change a single aspect of my thesis. I would do it over and over again, taking the bad with the good without a grain of hesitation. Thank you all for making my PhD so special!

Mes premiers remerciements vont au triplé de directeurs de thèse le plus improbable : Patricia la matheuse, Fabien le psychologue et Thomas le statisticien. Merci, Patricia, de t'être lancée avec moi dans ce fou mais merveilleux projet de thèse. Même les moments où tu as douté de moi ont été tellement précieux dans la construction de mon autonomie scientifique et personnelle ! Tu auras toujours une place spéciale dans mon cœur. Merci, Fabien, pour nos discussions intéressantes. Même si rares, elles m'ont aidée à affiner ma réflection et à construire mon identité de psychologue. Merci également de m'avoir fait comprendre l'importance de passer du temps sur la bibliographie. Merci, Thomas, pour ton précieux travail de médiation. Les mathématiques et la psychologie sont deux langages différents et cela amène des fois à des incompréhensions. Merci de m'avoir aidée dans cet ingrat mais fondamental travail de traduction et de patience. Merci également pour ton investissement scientifique et émotif.

I thank Royce and Susanne for accepting to review and evaluate my manuscript. Merci, Royce, d'avoir pris le temps de lire attentivement mon manuscrit. Merci pour tes copieuses appréciations scientifiques et tes questions intéressantes. Merci également d'avoir remarqué le fin travail des petits détails ainsi que la passion que j'ai mis dans l'écriture. Je me souviendrai toujours du moment où j'ai lu ton merveilleux rapport et de mon incommensurable joie. Thank you, Susanne, for your sound and scientific feedback. I

will never forget your warm compliment: "Every researcher working with quantitative evaluations of these types of experiments should read this thesis!".

Thank to Robert Goldstone, Pierre Pudlo, and Olivier Renaud for accepting to be a member of my PhD thesis committee. A special thank to Robert for his kindness and his scientific work. It is a true honor to have you in my commitee!

Ringrazio tutta la mia famiglia che mi ha accompagnato in questa avventura francese piena di ostacoli e forti emozioni. Grazie Mamy, Papi, Topis e Gabri di essere il mio punto fisso e di avermi permesso di volare per realizzare i miei sogni. Un grazie anche ai miei nonni Graziella e Nino per la loro costante presenza.

Un ringraziamento speciale anche a Dario, Eugenia, Primo, Simona e Tommaso per essere la mia seconda famiglia. Grazie, specialemente a te, Dario, per il tuo cuore grande e generoso.

Je remercie ma troisième famille, Anne, Bruno, Julie, Émilie, Trixy, Willow et Elios. C'est toujours une fête et une immense joie de venir vous voir. Merci de me faire sentir un membre de la famille, c'est le cadeau le plus beau que vous auriez pu me faire !

Merci aux doctorant.e.s, post-doctorant.e.s et stagiaires du laboratoire Dieudonné, BCL, I3S, INRIA et autres. Je n'ai pas eu l'occasion de lier avec tout le monde, néanmoins je tiens à vous remercier pour avoir rendu mon expérience de thèse plus agréable. Du laboratoire Dieudonné, je remercie Arthur, Edouard, Julie, Luis, Marcella, Reine, Pupi, Victor, Eliot, Alexis G., Huda, Sofiane, Ludovick, Marco, Maxime, Jonathan, Yash, Mehdi, Gaetan, Kevin, Hadrien, Léo, Billel, Nadine, Zhixin, Djaffar, Cuong, Valerio, Felice, Laurent, Giulia et Riccardo. Un remerciement particulier va à Cécile et Wesley pour avoir toujours accepté avec enthousiasme mes idées de sortie de la dernière minute. Je remercie également Sara pour nos discussions dans le couloir et pour valoriser mes conseils. Du laboratoire BCL, je remercie Sophie, Miriam, Raphaëlle, Camille, Moustapha, Samaneh, Laura, Alex, Paolo et Francesco. Du laboratoire I3S, INRIA et autres, je remercie Eman, Yuri et Ali.

Je ne peux pas ne pas remercier mes ex co-bureaux, Victor, Pupi et Sofiane pour avoir partagé avec moi les moments les plus difficiles de ma thèse.

Un grand merci aussi à tout le cinquième étage du bâtiment Fizeau : Cédric, François, Yves, Gilles, Roland, Rémi, Didier, Claire, Julie, Raphaël, Olivier, André, Luc, Damien, Elisabeth, Elena, Ivan, Marjolaine, Charles, Stella, Seb. Merci d'avoir su créer une ambiance très agréable et joviale. En particulier, je remercie Anna et Dimitri d'avoir partagé avec moi l'étage de Fizeau pendant de nombreuses soirées. Merci également, Anna, de proposer

souvent des activités de détente à tout l'étage ! Un merci spécial également à Martine, pour ton humour et nos nombreuses discussions autour de l'enseignement.

Je souhaite aussi remercier Indira pour avoir organisé le repas à l'occasion de la journée "Women in maths", Francesca pour ses invitations aux concerts, et Thierry pour sa bonne humeur.

Merci au meilleur couple d'informaticiens, Jean-Marc et Roland. Un merci très spécial à Jean-Marc pour nos discussions, sa disponibilité infinie et son constant soutien !

Je tiens également à remercier tout le personnel administratif pour leur travail indispensable : Amandine, qui se souviendra de moi pour les obstacles rencontrés dans l'achat des crédits PsychoPy, Narymen, les deux Isabelle, Chiara et Anita, sans oublié Julia qui m'a accueilli avec gentillesse à mon arrivée au labo.

Un merci très spécial à Marc. Avoir collaboré avec toi pour la création de jeux et activités pour la fête de la science a été une des expériences les plus enrichissantes et amusantes de mon doctorat. Merci pour ton précieux travail de diffusion scientifique qui permet de réunir les gens autour de la science. Merci également d'avoir accepté de m'aider dans la création d'une version interactive de ma thèse. J'espère pouvoir bientôt entreprendre une nouvelle collaboration créative avec toi !

Merci à ceux avec qui j'ai partagé l'organisation et l'animation de la fête de la science : Marc, Jean-Louis, Robert, Cyril, Ludovic, Maxime I., Jean-Baptiste, Jonathan, Eliot.

Merci à toutes les personnes avec qui j'ai partagé le plaisir d'enseigner. Merci à Ann pour sa patience et bienveillance. Merci à Nicole, Mohammed et Ingo pour avoir rendu ma première expérience d'enseignement très agréable. Merci à Raphaël, Nahla et Yash, avec lesquels j'ai passé une amusante journée et soirée à scanner des examens. Merci à tou.te.s mes étudiant.e.s pour les gentils et encourageants retours ! Merci également à toutes les personnes qui ont pris le temps de faire mon expérience psychologique.

Merci à toutes les personnes de l'institut NeuroMod. En particulier, merci à Alexandre et Indrig pour la belle et surprenante collaboration, et à Chloé pour son travail d'information et d'organisation. Je remercie également Tobias et Émilie pour les échanges lors des rencontres C@UCA et NeuroMod.

Quand je pense aux expériences clés qui m'ont amené à terminer positivement mon doctorat, la formation de co-orientation fait sûrement partie de cette catégorie. Un

grand merci à Catherine, Emmanuelle, Laurie et Isabelle d'avoir rendu cette formation mémorable. Grâce à vous chaque session était un pur plaisir !

A big thank to Marya for helping me gaining confidence in my speaking ability. Also, thank you for all of our interesting conversations about teaching and learning. I cannot wait to celebrate my PhD with you in front of a good cake!

Merci à Alice et à toute la bande pour l'organisation des pique-niques bien-être. C'était toujours très agréable de discuter avec des personnes ayant des formations très variées autour d'un bon vin et de la bonne nourriture.

Merci à Peggy de m'avoir offert des nombreux moments de détente. Sans toi cela aurait été beaucoup plus difficile de gérer le stress et les imprévus de ce doctorat.

Merci à David de m'avoir permis de reprendre contact avec mon côté musical. Merci également à Anne et Hélène d'avoir partagé avec moi l'apprentissage de la guitare.

Je souhaite remercier le centre LGBTQIA+ de Nice pour les samedis après-midi passés ensemble. En particulier, merci à Sarah d'avoir eu un rôle important dans la découverte de moi-même.

Merci à toute l'équipe de foot de Villeneuve-Loubet. Malgré les problèmes de santé et les nombreux accidents (côte, cheville, côte à nouveau) cela a été un rêve de pouvoir jouer dans une vraie équipe.

Merci à Nathalie, Stéphanie et Patricia de m'avoir initiée à la sophrologie dans un moment particulièrement difficile de ma thèse.

Je vais également me souvenir de ce doctorat pour ses ponctuels mais marquants voyages.

Merci à Cycy, Oussama, Anne et Yuri avec lesquels j'ai vécu la meilleure vacance de groupe de ma vie.

Un grand merci à Ophélie et Cycy pour l'inoubliable summer school à Bornholm. Je n'oublierai jamais nos soirées d'exploration, le séchage de quelques cours (désolée Patou ;-P), la bouffe et les grasses rigolades. Merci pour ce beau souvenir !

Merci à Miriam d'avoir partagé avec moi les jours à Pavia à l'occasion du symposium de psychologie. La visite de la ville ainsi que les conférences ont été beaucoup plus agréables grâce à ta présence.

Un grazie a Anna, Luna, Olaia e Francesco che hanno reso divertentissime le due vacanze solidali che ho passato alla Lipu.

J'aimerais remercier tous les restaurants niçois qui ont nourri mon coeur autant que mon estomac pendant ces quatre années de thèse. Merci à Koko Green de m'avoir fait découvrir une fine gastronomie végane, merci à Au Petit Libanais d'avoir nourri mes amitiés, merci à Pizz'Athena de m'avoir fait redécouvrir le goût de la pizza italienne en France, merci au Fairy Sushi d'avoir été le lieu de célébration de mes victoires, merci à l'Union d'avoir accueilli des nombreux repas entre doctorants, merci au Viking Burger qui a représenté un des petits plaisirs du confinement, et enfin merci au King Sushi qui m'a permis de gagner le cœur de ma princesse.

Enfin je souhaite remercier tou.te.s les ami.e.s qui ont illuminé, de prêt ou de loin, mes journées niçoises.

Merci, Laurence, pour ton soutien dans les moments difficiles ainsi que pour nos discussions stimulantes.

Merci, Célia, pour ton sourire, ta franchise et nos dîners ensemble.

Merci, Fatat, pour ta joie contagieuse et les week-ends à Paris. Merci de faire partie de ma vie malgré la distance qui nous sépare.

Grazie, Alessandra, per mantenere viva la nostra amicizia nonostante la distanza e per essere fedele a te stessa. Dà speranza vedere persone che creano una vita su misura senza seguire la massa.

Grazie, Elisabetta e Arthur, per i molteplici mercoledì sera passati a giocare e a discutere (quando Arthur ci dava un momento di tregua). Spero con tutto il cuore di poter seguire Arthur nella crescita ed essere per lui un ulteriore punto di riferimento.

Merci à Anne pour nos midis à L'Altra Casa. Je chérie chacune des nos discussions. Merci d'être aussi ouverte, courageuse et curieuse, c'est inspirant de te côtoyer.

Merci, Bochra, de m'avoir fait découvrir la culture tunisienne ainsi que d'avoir accueilli avec ouverture d'esprit mes nombreuses confidences. Merci également pour l'enrichissant échange de livres et pour nos agréables balades lors du confinement.

Merci, Oussama, pour les voyages à Pisa, à Vienne et chez les parents de Cycy, pour nos discussions d'économie et d'écologie, pour les excellents thés à la menthe marocains, pour les tajines et enfin pour ta joie de vivre et ton aura positive.

Merci, Samira, de m'avoir accompagnée dans cette aventure. J'ai eu une chance énorme d'avoir eu la possibilité de partager la chambre avec toi lors du colloque des doctorants de 2017. Tu ne sais pas combien les discussions que nous avons eu à cette occasion m'ont aidée à réaliser que je n'étais pas la seule doctorante à avoir des doutes et des problèmes d'encadrement. Merci d'avoir été là pour moi !

Et enfin merci à ma merveilleuse princesse Cycy, sans laquelle je n'aurais jamais pu terminer ce doctorat avec la même force, la même détermination et la même confiance. Tu m'as donné une maison sur laquelle m'appuyer, tu m'as permis de découvrir mon exceptionalité, tu m'as fait sentir entièrement aimée (défauts et manies comprises). Je n'oublierai jamais le vendredi matin qui a suivi la réception du rapport de Royce. Nous étions en train de nous préparer pour aller au labo et tu as décidé de mettre la chanson qui a accompagné ma remonté. Envahie par les souvenirs et les émotions, j'ai fondue en larmes. La chanson a fait remonter à la surface toute la souffrance liée à ce doctorat, l'exténuant travail, les sacrifices, mais aussi la fierté et la profonde joie pour tout ce que j'avais accompli. Merci pour ce moment magique. Je t'aime.

Cela serait un pêché capitale de ne pas partager avec vous la colonne sonore qui a accompagnée ma thèse. Je vous conseille de lire les paroles en écoutant la musique ;-) (Try Everything de Shakira).

Oh, oh, oh, oh, oh ($\times 4$)

I messed up tonight
I lost another fight
Lost to myself, but I'll just start again

I keep falling down
I keep on hitting the ground
But I always get up now to see what's next

Birds don't just fly
They fall down and get up
Nobody learns without getting it wrong

I won't give up
No, I won't give in till I reach the end
And then I'll start again
No, I won't leave
I want to try everything
I want to try even though I could fail

I won't give up
No, I won't give in till I reach the end
Then I'll start again
No, I won't leave
I want to try everything
I want to try even though I could fail

Oh, oh, oh, oh, oh
Try everything
Oh, oh, oh, oh, oh
Try everything
Oh, oh, oh, oh, oh
Try everything
Oh, oh, oh, oh, oh

Look how far you've come
You filled your heart with love
Baby, you've done enough
Take a deep breath

Don't beat yourself up
No need to run so fast
Sometimes we come last, but we did our best

I won't give up
No, I won't give in till I reach the end
And then I'll start again
No, I won't leave
I want to try everything
I want to try even though I could fail

I won't give up
No, I won't give in till I reach the end
Then I'll start again
No, I won't leave
I want to try everything
I want to try even though I could fail

I'll keep on making those new mistakes
I'll keep on making them every day
Those new mistakes

Oh, oh, oh, oh, oh
Try everything
Oh, oh, oh, oh, oh
Try everything
Oh, oh, oh, oh, oh
Try everything
Oh, oh, oh, oh, oh
Try everything

Try Everything, Shakira

Contents

Preface

The present thesis is highly interdisciplinary. Thus, I attempted to write a manuscript that is accessible to the widest range of people (psychologists, mathematicians, statisticians, pedagogists, etc.). Moreover, because of my love for pedagogy, I tried to convey ideas, concepts, and messages in the simplest terms.

Almost every concept is explained using three complementary ways of integrating information: words, images, and mathematical equations. The goal was to allow readers to integrate content using their most suitable learning style.

Two types of text boxes are disseminated all along the manuscript: a "dig deeper" text box that provides further information about the studied topics (denoted as BOX), and a "summary" box that summarizes the content of a specific section or chapter (denoted as TO SUM UP). The rationale was to target three different types of people: the regular person who will read the manuscript skipping most of the text boxes, the nerd who is so curious that won't miss a single word (text boxes included), and the busy person who will only read the summary boxes.

Whatever kind of person you are (regular, nerd, or busy) I hope you will enjoy the reading.

1

Introduction

Contents

Categorization is the ability shared by both human and nonhuman animals to separate objects into classes. This skill enables individuals to survive. For instance, human beings have the ability to recognize facial expressions of anger or threat rapidly and efficiently. This fast detection of angry faces allows quick response and thus represents a clear evolutionary advantage [Fox+00; GW06]. In the animal kingdom, vervet monkeys have developed different alarm calls to warn other group members about the presence of a predator (e.g., eagles, leopards or snakes). This early multiform warning system gives them a greater chance to escape thus reducing their mortality [CS90].

Outline of this chapter

This chapter provides an introduction to categorization, the domain in which the present thesis is grounded. Firstly, we offer basic definitions in this field of study and we present the canonical categorization experiment. Secondly, we introduce different types of presentation orders of the to-be-categorized stimuli (in the context of a categorization experiment) and review the literature on the topic. Then, we present formal models in psychology by providing the reader with an overview of the most famous categorization models and explain why their use should be more widely promoted. Finally, we present the scope, aim, and organization of the present thesis.

1.1 Categorization

Let us first give a few definitions. By *categorization* we refer to the process (or processes) of organizing entities such as objects, events, ideas, etc. into groups. A set of entities that forms a group is a *category* and the members of a category are called *exemplars* or *items*. The mental representation of a category is defined as a *concept* [GKC12; MC98]. Take, for example, the cognitive process with which we would classify our neighbor's dog "Spike" as a "Dog". This cognitive process is an example of categorization. "Spike" the individual animal is an exemplar of the category "Dog" and the way the mind represents the category "Dog" is a concept.

Theories: Exemplars vs. Prototypes vs. Rules

Most research on categorization revolves around the way categories are mentally represented, which traditionally has fallen under three main proposals: exemplars, prototypes, and rules [GKC12; Kru05; MC98; PW11].

Exemplars. According to the exemplar theory, animals store objects they encounter separately as unique memory traces so as to categorize new items by comparing them to these previously traces called exemplars. If the new items are enough similar to the stored exemplars of a given category, then would tend to be spontaneously categorized in the same category with no need to be corrected by a feedback [Bro78; Est94; MS78; Nos84; Nos87]. For example, according to this theoretical framework,

human and nonhuman animals mentally apprehend the category of “mushrooms” by memorizing all exemplars of porcini, chanterelle, amanita, morel, shiitake (and so on) they come across. When they detect a new item, the more similar this item is to the stored exemplars of mushrooms, the higher the chance to categorize it as a “mushroom” (see Figure 1.1). Conversely, it could be mistaken as a rock or a piece of wood, etc.

Prototypes. The prototype theory is based on the idea that animals store a summary representation of the many exemplars that form a category, and this mental summary is called a prototype. The prototype expresses the most common features of the members of the category. The categorization of new items consists of comparing them to the prototype of the category. If the new items are similar enough to the prototype, then they are included in the category [MR81; Ros75; RM75]. For example, according to the prototype theory, animals would average all exemplars of porcini, chanterelle, amanita, morel, shiitake (and so on) they come across to form the prototype of the category of “mushrooms”. When they detect a new item, they classify it as a “mushroom” if it is similar enough to the stored prototype (see Figure 1.1).

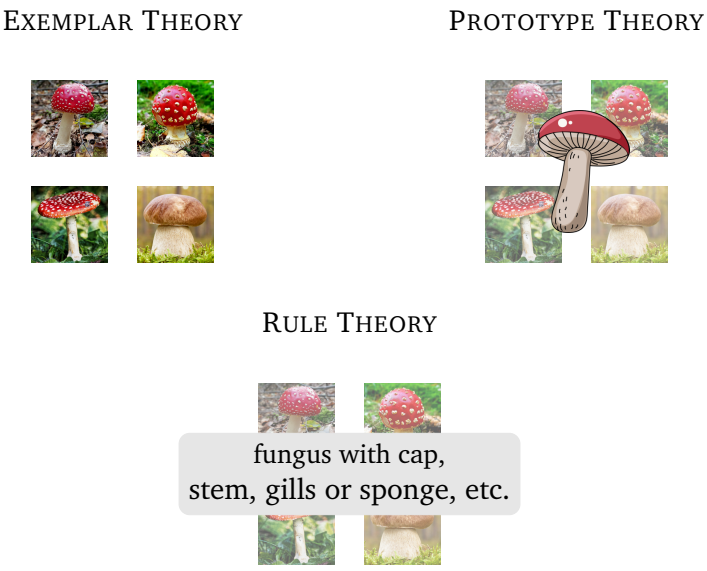


Figure 1.1 – *Example of exemplar, prototype, and rule theory. The image of the prototype “mushroom” has been taken from* `https://creazilla.com` *(Creazilla Open-Source License), whereas all the other images have been taken from* `https://pixabay.com/fr/photos` *(Pixabay Licence).*

Rules. According to the rule theory, animals store lists of necessary and sufficient features for category membership, called rules. New items are categorized as members of a category if they satisfy the rules of the category [BGA58]. For example, a "mushroom" is anything that is an eukaryotic organism of the group fungi, characterized by a cap, a stem, gills or sponge under the cap, etc. (see Figure 1.1). Rules can be updated and become more refined as objects become of particular interest for an individual, for instance when foraging for edible mushrooms.

The Role of Similarity

The intuitive idea that items are put into categories because they are perceived as similar has been quite influential. Indeed, many theories (as saw in the previous subsection) and categorization models (as we will see in Section 1.4) are grounded on similarity. For example, prototype theories assume that new items are classified into a category on the basis of their similarity to the category prototype, and exemplar theories assume individuals compute the similarities between the new item and all existing exemplars.

However, a number of theoretical arguments and empirical evidence have questioned the role of similarity in categorization [Bar82; Goo72; SOS92; ZH99]. Some researchers have argued that similarity is too unconstrained to serve as an explanation for categorization [Goo72; Rip89]. Others have argued that because similarity is largely based on perception it is difficult to take into account forms of abstractions that are frequent in concepts. Finally, rule-based theories assume that individuals rely on a set of abstracted features to judge whether a new object belongs to a category, which therefore cannot be referred to as a similarity process [Car85; Kom92; MM85]. Although it is clear that similarity is not sufficient to account for all categorization processes, it still does play an important role in establishing many of our categories [SSO93]. Excellent reviews on the role of similarity in categorization have been written by Goldstone [Gol94] and Medin, Goldstone, and Gentner [MGG93].

In the Present Thesis

The theories described in the present section (Section 1.1) also apply for non-human animals [AJT17; JOU11; Smi+16]. However, the present work focuses on humans, more specifically adults.

1.2 Canonical Categorization Experiment

This section is designed for those who are not familiar with categorization experiments. Below, we define the main features of a categorization experiment and introduce the proper vocabulary.

A Brief Example of Categorization Experiment

Imagine your neighbor Tom wants to learn about mushrooms. More specifically, he wants to acquire the ability to determine whether a mushroom is edible or not, which corresponds to the ability to categorize. One way to help him progress is to set up a categorization experiment. We could select a sample of pictures of edible and non-edible mushrooms (see Figure 1.2) and we would present these pictures to Tom, one at a time. For each picture, we would ask Tom to classify the mushroom into the edible versus non-edible category. After his response, a feedback would correct his answer. Hopefully, Tom would eventually learn the correct classification of all the selected pictures after viewing them multiple times. Usually, a similar process is at play with a friend in a forest, with real mushrooms. Furthermore, if Tom wished to test his general ability, we would select a set of new pictures and apply the same procedure, or more risky, Tom could take his chances and go directly in a forest pick what he thinks are edible mushrooms. The ultimate feedback would be if Tom remained alive or not.

The mushroom pictures that have been used to train Tom are called *learning items*, while those that have been used to test Tom's acquisition of knowledge are called *transfer items*. A single presentation of a mushroom picture is called a *trial* and the consecutive presentation of all the available learning (or transfer) items is called a *block* (see Figure 1.2). The mushroom picture presented at a trial is a *stimulus*.

Options in Categorization Experiments

The previous example gives a general idea about how a categorization experiment is conducted. However, those conducting the experiment have a larger array of choices than the one presented above. One example of particular interest in the present thesis is that instead of going in the forest to encounter mushrooms in a random way, the choice of mushrooms and the order in which they can be presented can be manipulated in a

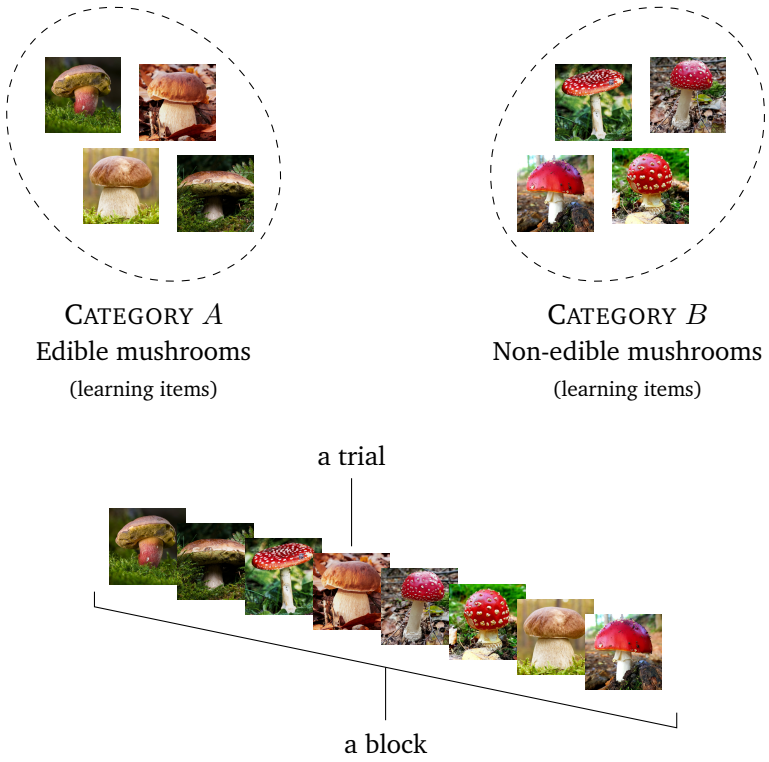


Figure 1.2 – *Example of a categorization experiment. On the top, the learning items are divided into two categories, respectively, edible mushrooms (category A) and non-edible mushrooms (category B). On the bottom, illustration of a trial (i.e., a single presentation of a mushroom picture) and a block (i.e., the consecutive presentation of all the learning items). Images taken from* `https://pixabay.com/fr/photos` *(Pixabay Licence).*

laboratory experiment. Let us explore some degrees of freedom that researchers face when designing categorization experiments.

Phases

One of the first aspects that researchers have to deal with is determining the number and nature of the phases of the experiment. For example, the experiment administered to Tom had two phases: a *learning phase* in which Tom was trained, and a *transfer phase* in which Tom was tested. The number and nature of the phases of an experiment depend on the objective of the experimenter. If the experimenter aims to determine the optimum learning condition (among a selection), then it is appropriate to design a single learning

phase [Gag50; MF09; NP96]. In contrast, if the goal is to study how well participants transfer their knowledge to new stimuli, then a transfer phase is necessary [CG14b; KP12; MF16; Mea+17; Noh+16; Nos86].

The transfer phase tests the generalizability/transferability of knowledge. Without the use of new items in this phase, it is possible for an individual to learn how to classify items by rote, which is not the goal of a learning experience. If the individual succeeds in classifying red mushrooms and brown mushrooms as mushrooms and black birds and yellow birds as birds, we can only be sure that learning generalizes if the subject is tested with a red (or brown) bird, which determines whether the subject is capable of inductive inference.

Items

Another aspect is the choice of the items (learning items, transfer items, or both). There are three main questions that can guide the choices of an experimenter [AM98].

Artificial or real-world items? Most of the research on categorization has been conducted with artificial items, such as geometric shapes [Gag50; MF09; MF16; Noh+16; Nos86; NP96; SHJ61], blob figures [CG14b; CG14a], or simplified drawings [SCR10]. However, there has been an increasing interest in using real-world items, such as rocks [Mea+17; Miy+18; Nos+19], birds [Bir+12; KV18], faces [GLS01], or paintings [KP12; KB08]. Researchers that employ real-world items face the additional choice of using pictures versus physical samples [Mea+18].

Continuous or discrete dimensions? The *dimensions* of an item are the main features of the item. For example, the main features of a red square are color and shape. Experimenters can choose to use items that vary continuously along their dimensions (for example, straight lines varying in orientation) [Gol96; Noh+16; Nos86], or that only take a discrete number of values (for example, geometric shapes such as square, circle and triangle) [MF16; NP96]. A widely studied case among those belonging to the latter group is when dimensions can only take two values (i.e., binary-valued/Boolean dimensions) [MF09; SHJ61].

Separable or integral dimensions? A last aspect that experimenters can choose concerns the interaction between the dimensions of the items. Separable dimensions are easy to distinguish from one another. Conversely, integral dimensions are perceived as not distinguishable. For example, color and shape are separable dimensions,

while color and brightness are integral dimensions. Research on categorization has focused on both separable [Noh+16; Nos86; SHJ61] and integral [Nos87; NP96] dimensions, but the present study only focuses on separable dimensions, which are more appropriate for comparing exemplar-based theories and rule-based theories.

Categories

Another set of options concerns the construction of the studied categories. Three main category structures are largely used in the literature.

Rule-based category structure. The categories belonging to this group are created from simple rules. An example could be the "white versus black rule", in which the white items are classified into one category and the black ones are classified into the opposing category. In order to be successful, participants have to determine the relevant dimension (in this example, the color dimension) while ignoring the others. Examples of experiments manipulating these types of categories can be found in [BGA58; PFM09].

Information-integration category structure. In this category structure every dimension is partially informative. Participants are required to integrate the information offered by all of the dimensions in order to correctly classify the items. Examples of information-integration category structures can be found in [Car+16; Mad+10; SA08].

Rule-plus-exceptions category structure. This category structure is similar to the rule-based category structure with the difference that the rule contains some exceptions. An example is the rule "all white items plus the green square versus all black items plus the red square". Experiments manipulating these types of categories can be found in [Nos86; MF09; SHJ61]. The rule or structure description is most often based on one of the two categories, such as "all white except the square, plus the black square", meaning that all other objects go into the opposing category.

Feedback

After the items are selected and the categories constructed, the experimenter has to decide whether feedback should be provided to the participants or not. In a *supervised* task,

participants are informed about the correctness of their responses. In an *unsupervised* task, no feedback is given after each trial. An absence of feedback does not mean that the categorization is impossible. As seen earlier, a categorization experiment is most often composed of either one (learning phase) or two phases (learning and transfer phases). The learning phase is generally supervised while the transfer phase is generally unsupervised [CG14b; MF16; Nos86]. The rationale is that we wish to study a spontaneous classification in the transfer phase, so this phase should not be supervised.

Stop Criteria

There are two main stop criteria that the experimenter can adopt. The first option is to end the experiment (or a specific phase of the experiment) when the participant completes a fixed number of trials. The second option is to set a specific criterion based on the performance of the participant. When the participant reaches the criterion, the experiment (or a specific phase) ends. A plausible performance criterion could be to end the experiment once the participant gives m correct responses in a row. It is common to come across experiments that include both stop criteria. In these experiments, the first option is used during the transfer phase while the second option is used during the learning phase [MF16; Nos86]. There is no perfect choice of criterion for the learning phase. When the number of blocks is fixed, all participants receive an equal number of stimuli. In contrast, when we request a fixed rate of success, for instance 90%, participants are guaranteed to reach equal performance but can have different experience (e.g., some of them receive a greater number of stimuli). The disadvantage of using a fixed number of blocks is that some participants sometimes cannot achieve minimal performance.

Between-category presentation orders

Another important aspect that researchers have to deal with is the order in which categories are presented (the so-called *between-category* presentation order). There are three main between-category orders:

Interleaved. Interleaving means that the studied categories are presented alternately. For example, let us consider two categories (A and B). The categories are interleaved when each pair of adjacent stimuli belong to different categories, for instance, when categories alternate every trial as in $ABABABAB$. However, a strict alternation

would be nonsense since the participant could alternate the response keys to be 100% correct without even looking at the stimulus features. Interleaving is thus usually obtained by randomly permuting stimuli, so as to produce orders in which 70%, 80%, or 90% of consecutive pairs of stimuli belongs to opposing categories ($AABABBAB$; see Figure 1.3).

Blocked. Blocking means presenting members of the same category in successive trials. For example, the categories A and B are blocked when all the members of category A and B are grouped together. Similarly to interleaving, it would be nonsense to strictly block categories (again, the participant could block the response keys to be 100% correct without looking at the stimuli). Therefore, blocks in which categories are strictly blocked are generally alternated with random blocks. Another solution would be randomly permuting stimuli so as to produce orders in which 70%, 80%, or 90% of consecutive pairs of stimuli belongs to the same category ($AAABBBBA$; see Figure 1.3).

These types of between-category presentation orders have been used extensively by experimenters in the literature, and especially because there has been a long tradition of studies of blocked practice versus interleaved practice in the domain of learning in general [Bir+12; CG14b; CG14c; CG14a; CG15; CG17; Gag50; KP12; KB08; Kor+10; KV18; Noh+16].

Figure 1.3 – *Example of interleaved and blocked study (between-category orders). The learning items are the same as in Figure 1.2. Images taken from* `https://pixabay.com/fr/photos` *(Pixabay Licence).*

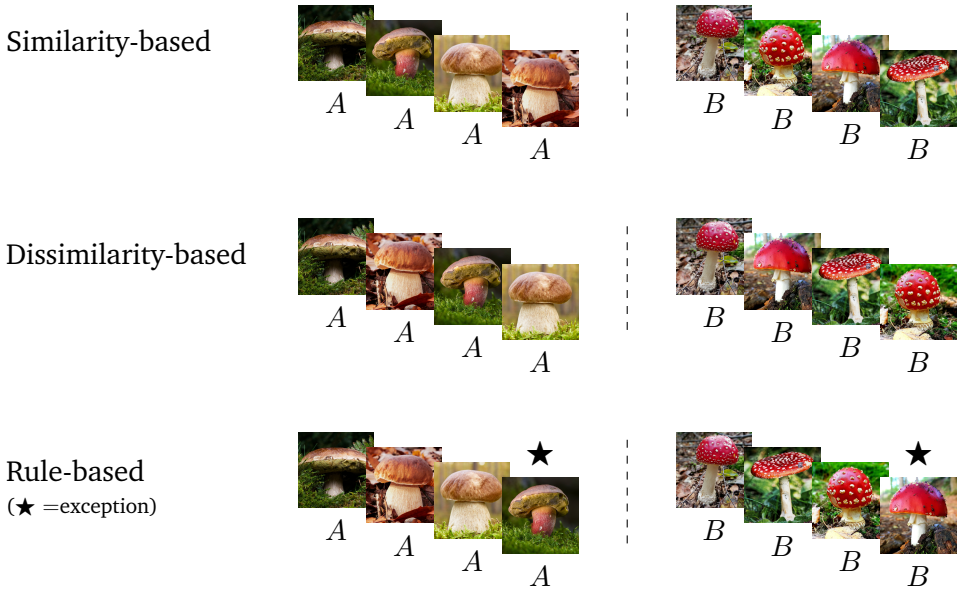


Figure 1.4 – *Example of similarity-based, dissimilarity-based, and rule-based study (within-category orders). In the rule-based order, exceptions are indicated with a star. The learning items are the same as in Figure 1.2. Images taken from* `https://pixabay.com/fr/photos` *(Pixabay Licence).*

Within-category presentation orders

After selecting the between-category order, the experimenter can decide to manipulate the order in which members within a category are presented. There are three main options:

Similarity-based. Members within a category are presented following a similarity-based order if the similarity between successive stimuli is maximized (see Figure 1.4).

Dissimilarity-based. Contrary to the similarity-based order, the dissimilarity-based order is designed to minimize the similarity between successive stimuli (see Figure 1.4).

Rule-based. According to the rule-based order, stimuli are presented following a "principal rule plus exceptions" structure. This means that the members that most represent the category are presented before those that are less typical. For example, let us consider the category of mushrooms. Exemplars of Boletus edulis are more representative of the category "mushrooms" than exemplars of Boletus erythropus. Therefore, in a rule-based order, exemplars of Boletus edulis are presented

strictly before exemplars of Boletus erythopus (see Figure 1.4). Another important component of this type of ordering is that it was originally designed to favor an abstraction process by randomly presenting the cluster of stimuli encompassed by a rule [MF09].

These types of within-category presentation manipulation of orders are quite rare in the categorization domain [EA81; EA84; MF09; MF16], but also in other context, such as word recall [Bow+69] and old-new recognition tasks [MB94]. One specific reason why these manipulations are rare is that researchers might have thought that specific orders could induce specific learning processes, whereas research most often attempts to describe general processes.

Orders across blocks

Once the between and within-category presentation orders have been chosen, the experimenter has still to decide whether presentation orders are constant across blocks. By using a *constant* presentation across blocks, all blocks would be identical, meaning that the same sequence of stimuli would be presented in the successive blocks. By using a *variable* presentation order across blocks the sequence of stimuli could vary from a block to another (see Figure 1.5), but obeying the constraints of the chosen options (for instance, a similarity-based order could start every new block with a randomly chosen first item and select successive items by maximizing their similarity to the previous adjacent item). The variable presentation across blocks can be considered as the default random presentation. The next section gives more details about how these types of orders have been studied in the literature.

1.3 Presentation Order

If the phrase “men eat apples” delivers a feeling of everyday situation, the phrase “apples eat men” sounds more like the next science-fiction movie or a good cartoon. Despite the fact that these phrases share the same words (presented in different orders), the message sent is completely different. When the syntax does not fit the right vocabulary, the correct meaning cannot be conveyed. Some of us may have experienced as a child the powerful effect of telling a bad grade right after a good one to our parents, rather than announcing the two grades in the reverse order. The widely known C major scale can communicate

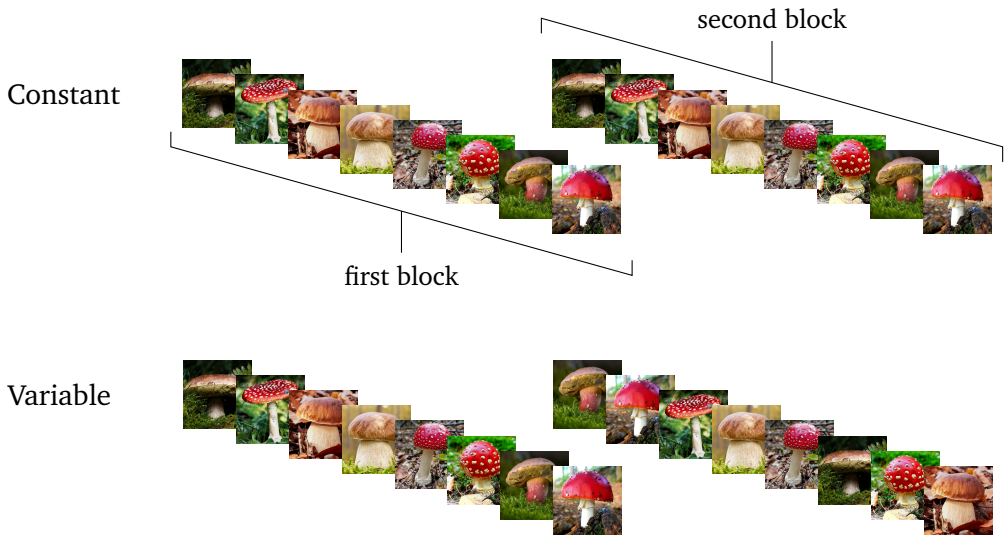


Figure 1.5 – *Example of a constant and variable presentation across blocks. The learning items are the same as in Figure 1.2. Images taken from* `https://pixabay.com/fr/photos` *(Pixabay Licence).*

to the audience a sense of joy or sadness, depending on the order in which its notes are played (do re mi fa sol la ti versus ti la sol fa mi re do, or equivalently, C D E F G A B versus B A G F E D C).

This intuition that the same content using different presentation orders can lead to distinct results has been confirmed by experimental data [BS81; Bra08; Cla14; Cor+11; EA84; HGV11; JS03; LCK12; Lip61; MP15; MFP13; QA14; Sam69; SD08; ZM09; ZJM11]. These studies have shown evidence that the order in which information is presented can impact the way we perceive, represent and learn new information.

As seen in the previous section, there are two main (interrelated) directions that researchers have taken in order to investigate presentation orders: by comparing interleaved versus blocked between-category orders and by studying the effect of similarity-based versus rule-based within-category orders. In addition, other studies have focused on more local effects of presentation orders such as determining how the previous stimulus affects the classification of the next stimulus. This section aims to review the literature on these topics.

Interleaved vs. Blocked Between-Category Orders

There is a considerable amount of research analyzing the differences between interleaving exemplars of different categories versus blocking members of a same category (see definitions for interleaving and blocking in Section 1.2) [Bir+12; CG14b; CG14c; KB08; Kor+10; KCG15; Roh09; Roh12; SYK16; Yan+17; ZB12; Zul+12]. Most research comparing these two conditions (interleaving versus blocking) concluded that interleaved study is more beneficial than blocked study [KP12; KB08; WFJ12; Zul+12].

For example, Kornell and Bjork [KB08] showed that participants who learned paintings from 12 different artists by interleaving paintings by different artists (rather than blocking paintings by artist) performed better at categorizing new items (Experiments 1a and 1b; see also [Kor+10]). Moreover, participants who followed the interleaved condition were also better at determining if a new painting was painted by a previously studied artist or a new artist (Experiment 2). Surprisingly though, the participants had the reverse feeling that the blocked order would lead to better performance because it gave them a sense of fluid learning (see also [Yan+17]). Interleaving therefore can appear to induce between-category comparisons, which can be more difficult than simply searching for similarities within a give category. Similar results have been found using different items or procedures [Bir+12; CG14b; LCK12; TR10; WDJ11; ZB12].

However, there is also evidence showing that blocked study is more beneficial than interleaved study. For example, Carpenter and Mueller [CM13] showed that non-French speakers performed better in learning orthographic-to-phonological mappings in French (i.e., "ou" and the corresponding sound [u] in the words "mouton", "bouton", "genou"; and "eux" and the corresponding sound [ø] in the words "paresseux", "osseux", "somptueux") when words with the same mapping were blocked rather than interleaved. This result has been replicated with different stimuli and procedures [CB12; CG11; CG14b; ZM12; Gol96; KH56; RTJ14; ZB12]. Although the amount of research supporting the benefit of blocked study is smaller than that supporting the benefit of interleaved study, these contrasting findings raise the question of which cognitive mechanisms predict the differences between interleaving versus blocking.

Temporal spacing (i.e., the temporal delay between repetitions of the same category, the so-called "spacing effect" in memory, [Ebb13]) was one of the first proposal aiming to account for order effects. Because of the beneficial effects of spacing [BS81; CD05; Cep+09; DVS10; Gle76; GL80; KB11; KR10; LB08; Pas+07; RLP08], some researchers

proposed to explain the advantage of the interleaved over blocked study in terms of temporal spacing [KB08; WDJ11]. When compared to blocked study, interleaving categories results in a higher temporal delay between repetitions of a same category. Therefore, interleaving might be more beneficial than blocking since it increases temporal spacing.

The study conducted by Kang and Pashler [KP12] aimed to directly test the possibility that the benefit of interleaved study is produced by a greater temporal spacing. In their study, they evaluated participants' test performance on three conditions: interleaved, blocked and temporal spaced. In the temporal spaced condition, repetitions of each category were spaced in time but not interleaved. The temporal spaced study featured the same temporal spacing between same artist paintings as in the interleaved condition, but blocking paintings by artist. For instance, the blocked, interleaved and temporal spaced conditions would be respectively $AAAABBBB$, $ABABABAB$ and $A-A-A-A-B-B-B-B$ (where "-" represents a blank screen and filling pictures). The results showed that participants in the interleaved condition achieved the best performance (during the transfer phase). Moreover, participants who followed the blocked and temporal spaced studies showed equivalent performance (again, during the transfer phase). Similar results were confirmed using different materials and tasks [Bir+12; MNH08; Noh+16].

These findings led to the discarding of the hypothesis according to which the benefit of interleaving is due to temporal spacing. Carvahlo and Goldstone proposed an alternative parsimonious theory: the Sequential Attention Theory (SAT) [CG14a; CG14b; CG14c; CG15]. The sequential attention theory hypothesizes that during category learning participants compare the current stimulus with the previous one and attend to similarities or differences between the two items, depending on their category assignment. This means that, if the previous and current stimuli belong to the same category, participants' attention will be directed toward their similarities. Inversely, if they belong to different categories, participants' attention will be directed toward their differences.

Recent studies have found evidence supporting the sequential attention theory. For example, Carvalho and Goldstone [CG14b] studied blocking versus interleaving with either low similarity or high similarity categories. They found that blocked study improved classification performance on new items for low similarity categories, whereas interleaved study improved classification performance on new items for high similarity categories. Similar results have been obtained with different items and procedures [CG14c; KCG15; RTJ14; ZB12]. It is also important to underline that the sequential attention theory is consistent with previous studies that showed a recency bias (i.e., a cognitive bias that

favors recent events or stimuli over historic ones) during category learning [JLM06; JS03; SB04; SBC02; ZJM11].

Sequential Effects

Sequential effects refer to the influence of recent information (i.e., previous stimuli) on performance in repeated tasks. A significant amount of research has suggested that sequential context plays an important role in influencing participants' judgment [Bou93; Gar53; Mur62; Mye76; TW84]. For example, in absolute identification tasks (in which participants are asked to label stimuli in a one-to-one mapping), it has been found that current stimuli are judged as more similar to previous stimuli than they actually are [Gar53; HL68; Lac97; Mor89; WL70].

This phenomenon, called *assimilation effect*, has on opposite equivalent called *contrast effect*. According to the contrast effect, current stimuli are judged as more dissimilar to previous stimuli than they actually are. Contrast effects have also been observed in absolute identification tasks [HL68; Lac97; WL70]. For example, Holland and Lockhead [HL68] asked participants to label 10 loudness stimuli with numbers 1 through 10. The results showed an assimilation effect between the stimulus of the current trial and the immediately preceding trial. Moreover, a contrast effect has been found between the current stimulus and the preceding two-to-five trials.

Although assimilation and contrast effects are well documented in absolute identification tasks, they have received less attention in the categorization literature. A few studies have shown evidence for a contrast bias in categorization tasks [JLM06; SB04; SBC02]. For example, in the study conducted by Stewart et al. [SBC02], participants were asked to classify 10 equally spaced tones into two categories, with the five lowest tones as members of category A and the five highest tones as members of category B. They found that categorization of the current stimuli was more accurate when preceded by a distant member of the opposite category than by a distant member of the same category. Moreover, in a successive study, Stewart and Brown [SB04] observed that a contrast effect could also be produced by stimuli presented two trials back.

If contrast effects received modest attention in categorization, assimilation effects received an even more limited interest. Only few studies have shown evidence of assimilation effects [HY12; ZJM11]. For example, Hsu and Yang [HY12] conducted a categorization task involving 10 expressions from the fear-disgust continuum (the expressions ranged

from the prototypical expression of fear to the prototypical expression of disgust in 10 steps). They found that assimilation effects occurred when the distance between the preceding and current stimuli was small, while contrast effects occurred when the distance between the preceding and current stimuli was high.

Some proposals have attempted to explain the contrast and assimilation effects in categorization. For example, Stewart et al. [SBC02] argued that the same-category contrast in classification is generated by an estimation process called the Memory and Contrast (MAC) strategy. According to the MAC strategy, participants estimate the relative difference between successive stimuli. If the lapse of time between the current stimulus and the one at the immediately preceding trial is high, then the participant tends to classify the current stimulus into the opposing category. Although the MAC strategy provides a good explanation for the same-category contrast, it fails to reconcile both the same-category contrast and the different-category assimilation.

In contrast, Jones et al. [JLM06] proposed a plausible solution for explaining both sequence effects. Their model of sequence effects in category learning (SECL) assumes that classification decisions are guided by two distinct mechanisms: decisional recency, which is the effect of the previous response, and perceptual recency, which is the effect of the previous stimulus. The notion of perceptual and decisional recency as two independent effects reconciles findings for both same-category contrast and different-category assimilation. The same-category adjustment can be viewed as a perceptual process in which, in the absence of the stimulus' information in classification, subjects use the location of the item presented on the preceding trial to fine-tune the representation of the category. Since participants tend to generalize their responses to the category they chose on a preceding trial, then cross-category effects could be the result of decisional recency. However, note that although the SECL theory offers a sound explanation for both sequence effects, it cannot be generalized to multiple category design (see [ZJM11]).

Similarity-Based vs. Rule-Based Within-Category Orders

Contrary to the between-category orders, which have been largely studied in the literature (see paragraph entitled "Interleaved vs. Blocked Between-Category Orders"), within-category orders have received more modest interest. Originally investigated in word recall tasks [Bow+69], the study of within-category orders has rarely been extended to

categorization tasks [EA81; EA84; MF09; MF16] and old-new recognition tasks [MB94] over the last few decades.

In categorization, research has recently been focused on contrasting similarity-based versus rule-based orders. In the similarity-based order, stimuli (of a same category) are arranged in order to maximize the similarity between adjacent exemplars, while in the rule-based order, the most representative exemplars of the category are displayed before the less typical exemplars. The first results comparing these two conditions concluded that rule-based order is more beneficial than similarity-based order (during both learning [MF09] and transfer [EA84]).

For example, Mathy and Feldman [MF09] found that participants learned to classify 16 stimuli - of varying shape, color, size and filling pattern - the fastest when they received a rule-based training as compared to a similarity-based training (see also [MF16] for similar results). Additionally, they observed that participants who followed either a similarity-based or a rule-based order performed better than those who followed a dissimilarity-based order (in which stimuli of the same category were arranged in order to minimize the similarity between adjacent exemplars). Note that the choice of categories used in these studies might have favored the rule-based order (in particular because the structures were logical by nature), but the goal of the present thesis is rather to account for presentation orders in general, rather than searching to increase categorization performance by finding the most optimal presentation order.

The study of similarity-based versus rule-based order is particularly relevant since the creation of these two types of presentation orders has been inspired by two contrasting ways of learning: a process based on associative mechanisms and an inductive process based on abstraction [Slo96]. The similarity-based condition follows an associative process that uses the temporal proximity of the stimuli to strengthen the memory traces of the two stimuli and, by extension, the entire similarity structure. In contrast, the rule-based condition aims to induce participants to form a logical rule. Grouping the most representative exemplars should help learners abstract a simple logic describing the stimuli.

For example, Mathy and Feldman [MF16] investigated whether the order that participants received during learning (similarity-based versus rule-based) shaped their mental representation of the categories. Their results showed that participants who received a rule-based training exhibited generalization patterns (i.e., the categorization of new items) consistent with rule-based retrieval. However, although participants who received

a similarity-based training showed distinct generalization patterns, these generalization patterns were not altogether consistent with exemplar-based retrieval.

Moreover, research about similarity-based versus rule-based order might be crucial for testing rule- and exemplar-based models of categorization [MS78; Nos86; RR04], as well as incremental-learning models [GB88b; Kru92; LMG04; SJL08]. Effectively, ruled-based models should be more sensitive to a rule-based presentation than to a similarity-based presentation and vice-versa for similarity-based models. The idea is that a model built to form a certain type of representation should lean toward an order that fits the shape of the target representation.

The current research regarding the study of within-category orders in categorization has been limited to particular contexts. For example, in [MF09] the presentation of the stimuli of different categories was randomized (i.e., random between-category order). To the best of our knowledge, only one study [MF16] investigated similarity-based versus rule-based order when categories were blocked (blocked between-category order). Therefore, a specific aim of the present thesis is to investigate whether rule-based training facilitates learning in different contexts, such as with an interleaved between-category order (the aim will be presented in detail in Section 1.5).

ON THE PRESENT THESIS

One of the main focuses of the present work is on investigating the effects of similarity-based versus rule-based within-category orders. We expect these effects to be modulated by between-category orders, for instance when categories are blocked instead of being interleaved.

1.4 Principles of Cognitive Modeling

Cognitive models are powerful tools for simulating cognitive processes in human and nonhuman animals. The approximation of cognitive processes allows researchers to test cognitive hypotheses and predict behavior. The great benefit of cognitive models to fit quantitative data explains their frequent use in cognitive science, for instance in domains such as categorization, memory, but also decision making [CG19; HK01; Hsu+19; NSM17; NSM18a; Nos+18; RH05; RR04; SN20; SM00].

A Brief Example of Cognitive Model

An effective way to understand how a cognitive model functions is to create one from scratch. Let us suppose that someone who loves to memorize new words wishes to investigate why some words are memorized easier than others. A first step is to make some plausible hypotheses about the hidden cognitive mechanisms at play.

We notice that long words are generally harder to recall than short words. Therefore, we conclude that the hypothesis "the shorter the word, the easier the storage" might be a good one to test. In order to create a cognitive model from this hypothesis, we need to reformulate the verbal statement in mathematical terms. For instance: "the ease of learning a new word is inversely proportional to its length", which is equivalent to the following equation:

$$e(x) = \frac{k}{l(x)}, \tag{1.1}$$

where x denotes a new word, e the ease with which a word is stored, l the length of the word, and k a freely estimated parameter to tune the model a bit in order to best fit the data (which can help produce more plausible predictions).

Equation 1.1 represents the core of our new cognitive model. Once the cognitive model is created, it has to be tested on real data. This allows one to evaluate the underlying hypothesis. If the predictions of the model are close to the experimental data, then the hypothesis is consistent with the real data. If not, the theoretical framework underlying the model needs to be reformulated. Model development is a never-ending process because virtually any model can be falsified, therefore more accurate ones are constantly being developed [Pop59].

What Is a Cognitive Model?

Cognitive science aims to understand how the brain accomplishes complex tasks, such as learning, remembering, predicting, and problem solving. The goal of a cognitive model is to account for one or more of these cognitive processes and clarify their respective interactions [BD10].

Computational cognitive models are generally described using rigorous mathematical or computer languages [BD10]. This aspect differentiates them from ungrounded conceptual models, or theories, which are instead based on verbal language which can often be

imprecise and result in fallacies or misinterpretations. In contrast, cognitive models make explicit and formal hypotheses by means of algorithms or equations. For example, Craik and Lockhart's "levels of processing" hypothesis [CL72] offers a theoretical framework for memory, whereas Shriffrin and Steyvers's model of recognition memory (REM) [SS97] expresses in mathematical terms how information is retrieved from memory.

A hallmark of cognitive models is that *they are based on cognitive principles* [BD10]. This is what makes cognitive models different from generic statistical models (or empirical curve-fitting models). For example, regression models or time series models can be applied to any type of data, as long as those data satisfy the statistical assumptions of the models (such as linearity or Gaussianity), whereas a cognitive model would not work to predict weather.

Cognitive models are also different from neural models. Neural models describe information processing at a more fine grain level, to describe the activity of the neurons as well as their interactions. On the contrary, *cognitive models explain human behavior at a more abstract level* [BD10]. For example, the Hodgkin-Huxley model [HH52] describes how action potentials in neurons are initiated and propagated by modeling the mechanisms involving ion channels in the neuronal membrane. In contrast, connectionist models (see [GB88b; Kru92]) do not aim to describe precise neural mechanisms, although they are inspired by the neural organization.

Why Would One Use Cognitive Models?

Here we attempt to show the potential as well as the contribution of cognitive models to psychology. There is a wide range of methods that can be used to study concepts and categories, for instance, designing a categorization experiment, creating a cognitive model, providing a theoretical framework, or even carrying out observational studies. All methods seem essential in the quest to shed light on the diverse aspects linked to the processes underlying categorization.

A first advantage of cognitive models over other approaches is that a cognitive model is described in rigorous mathematical language (see paragraph entitled "What Is a Cognitive Model?"). *The fact that the model has to be expressed in mathematical terms forces researchers to be explicit about their hypotheses and theories* [BD10; Mur11]. For example, if a categorization theory states that the probability of classifying two items into the same category is directly proportional to their similarity, this statement relies

on the reader's intuitive sense of similarity. It is only when we decide to implement a model based on this statement that one can realize that our computer (although it is the best available on the market) does not have any intuition of what similarity can be in algorithmic terms. This processing difficulty leads researchers to explicitly define what the similarity between two items can be in computational terms. Many concepts seem intuitive and easy on a verbal level, but expressing these concepts algorithmically is generally challenging.

A second reason for using cognitive models is that *they are capable of generating precise quantitative predictions* [Nor05; Mur11]. Predictions can be used to validate, dismiss or modify the assumptions underlying the model.

A third benefit of adopting cognitive models is that *they help ensure reproducibility in cognitive science* [FL10]. By implementing a model as a set of equations or a computer program, any researcher can reproduce the findings of a study that uses cognitive models to generate predictions.

Finally, *cognitive models could help guide the search for effective learning strategies in education*, as proposed by Nosofsky et al. in [Nos+18]. As seen in Section 1.2, the choice of items, categories, presentation orders and so on is almost infinite. Therefore, cognitive models could be used as a tool for selecting the most promising teaching methods. Empirical tests would then be focused on testing the best strategies resulting from the model simulations.

How to Assess the Goodness of Fit of a Cognitive Model

There are various criteria that are used to assess the fit of the data provided by a model. We review the two most commonly used criteria: the likelihood and the least-square contrast [BD10].

Likelihood. The likelihood measures the probability that a model would have generated the observed data as a function of the parameters of the model [GV60]. The higher the likelihood, the higher the probability that the observed data have been generated by the model with the chosen set of parameters.

Least-square contrast. Perhaps the most commonly used method to assess the goodness of fit of a model is to sum the squared deviations between the observed and predicted values [Leg05]. The smaller the sum of squared deviations, the higher

the goodness of fit. In cognitive science, this method is most commonly called Sum of Squared Deviations (SSD), or Sum of Squared Differences (SSD), or Sum of Squared Errors (SSE), or the Residuals Sum of Squares (RSS). In the present work we adopt the term Sum of Squared Deviations (SSD).

The previously reviewed criteria can also be used to estimate the parameters of a model. For example, the set of parameter values that best fits the data can be found by maximizing the likelihood (as a function of the parameter values) or by minimizing the sum of squared deviations (again, as a function of the parameter values). To anticipate, in the present work, the likelihood is used to estimate the parameters of the models, while both the likelihood and sum of squared deviations are used to evaluate the fit of the model (details are given in Chapter 4).

Categorization Models

In this subsection we explain the most common way categorization models are generally grouped (for a more exhaustive description of how models can be grouped, see [Wil13] and [Kru08]).

Input Representation

All cognitive models make a set of assumptions about the nature of the information they receive. Regarding models of categorization, those assumptions typically take two forms: geometric or featural. In the *geometric* input representation, stimuli are represented as points in a psychological space and are expressed in terms of precise coordinates. Two stimuli are considered as similar whenever they are close to each other in this psychological space. In contrast, in the *featural* input representation, stimuli are represented as a set of features. Two stimuli are considered as similar when they have common features. For example, the Generalized Context Model (GCM) [Nos86] adopts the geometric input representation, whereas the Component-Cue model [GB88b] adopts a featural input representation (both models are studied in more depth in Chapter 3)

Mechanisms of Attention

Some models of categorization assume that the information they receive can be modulated by attentional mechanisms. The implementation of this cognitive function aims to improve categorization accuracy. Depending on the nature of the input representation (geometric or featural), attentional modulation operates either at the level of the dimensions of the psychological space [Nos86] or at the level of features [Kru01; Mac75]. At the level of dimensions, attentional mechanisms result in stretching and shrinking the psychological space, for instance by mostly paying attention to the size of an object at the expense of other features (this point is clarified in Chapter 3). At the level of features, attention to a specific feature expresses the fact that the feature is a good predictor of the category membership (for instance, if a given size is diagnostic of a malignant tumor).

Some Non-Exclusive Classes of Categorization Models

In this paragraph, we present a non-exhaustive list of classes of category learning models which are rooted in theories about the mental representation based on exemplars, prototypes, and rules; see Section 1.1).

Exemplar models. Exemplar models store every distinct occurrence of an item (as well as its category membership) and classify a new item as a function of its similarity to all of the previously stored items. Nosofsky's Generalized Context Model (GCM) [Nos86], Medin and Schaffer's context model [MS78], Estes's array-similarity model [Est86], Nosofsky and Palmieri's EBRW model [NP15], and Kruschke's ALCOVE model [Kru92] are examples of exemplar models.

Prototype models. A prototype model operates in the same way as an exemplar model, but instead of storing every encountered item, it only stores a summary representation (called the prototype) of the many items representative of a category. The category membership of a new item is determined by its similarity to those stored prototypes. Examples of prototype models include the Reed's comparative distance model [Ree72] and the Massaro and Friedman's Fuzzy Logical Model of Perception (FLMP) [MF90]. Prototype models can appear less computationally demanding than exemplar models, but also appear to reflect cognitive processes found in experimental results [RM75].

Rule models. Similar to prototype models, rule-based models specify categories by a summary of their content (i.e., a rule). A rule is a list of necessary and sufficient features that define category membership. Some of the best known rule-based models are Levine's hypothesis-testing approach [Lev75] and Nosofsky's RULEX [NPM94; NP98]. The latter also has the ability to memorize individual items (exceptions) in addition to rules. Other rule-based models have focused on minimization theories to account for the observation that the subjective difficulty of a categorization task is proportional to the logical incompressibility of the rule [Fel00; SHJ61].

Connectionist models. Connectionist models (also called connectionist networks) are based on Artificial Neural Network (ANN) [Dre90; Ros58; WJ74] and are inspired by biological neural networks. They implement an error-driven mechanism (also named back propagation or gradient descent mechanisms) allowing them to learn through trial and error. Examples of connectionist models include Kruschke's ALCOVE model [Kru92], Kruschke and Johansen's RASHNL [KJ99], Love, Medin and Gureckis's SUSTAIN [LMG04], Gluck and Bower's Component-Cue [GB88b], and the Configural-Cue model [GBH89]. Connectionist models can belong to one of the classes previously listed (exemplar, prototype, and rule models). For example, ALCOVE is grounded on the exemplar theory, whereas the Component-Cue model is grounded on the rule theory.

Adapting clustering models. The three options that concern the nature of category representations (exemplars, prototypes, and rules) are static by nature (i.e., they give an idea of the average difficulty of a categorization task, but they cannot describe the learning process when they are not implemented, for instance in a neural network). Adapting clustering models represent an alternative to these fixed-representation accounts. According to adapting clustering models, human memory is an adaptive process which is capable of encoding both highly specific information and abstract generalizations. This flexibility is incorporated in the models by means of clusters, which represent particular subsets of similar items either in an abstract or a specific way. Examples of adapting clustering models include Anderson's rational model [And91], Pothos and Chater's Simplicity Model (SM) [PC02], and Love, Medin and Gureckis's SUSTAIN [LMG04].

Hybrid models. Hybrid models are those that use multiple representations. For example, the COVIS model [Ash+98] and ATRIUM [EK98] integrate in one system both rule-based and exemplar-based approaches. In contrast, Busemeyer, Dewey and

Medin [BDM84] and Smith and Minda [SM00] combined prototype and exemplar models.

Ability to Learn

Another way to group categorization models is to consider their ability/inability to take into account the temporal dynamics of the learning process. There are categorization models that are capable of reproducing a learning curve (like ALCOVE [Kru92]) and others that are not (like the Generalized Context Model [Nos86]). These two classes of models have highly different mathematical properties. Therefore, we believe that this new distinction could help researchers select the categorization model that best suits their purpose. In what follows, we first define transfer models (which do not account for the learning dynamics) and learning models (which can reproduce learning), and secondly, we provide a description of their way of utilization.

Definition 1.1. Let M be a model and θ its set of parameters. Let $\mathbb{P}_M^{\theta,t}(A \mid \xi)$ denote the probability of classifying the item ξ into category A at time t, given M and θ. Let us also assume that the presentation order is constant across blocks (this hypothesis allows us to include in the definition models that integrate stimuli manipulation). If we have that

$$\mathbb{P}_M^{\theta,t}(A \mid \xi) = \mathbb{P}_M^{\theta,s}(A \mid \xi),$$

for all s and $t \in \mathbb{R}$, then we say that M is a *transfer model*. Otherwise, we say that M is a *learning model*. ⊠

In other words, transfer models classify items into a specific category with the same probability at any given time of the learning process (mathematically, they are stationary models when the presentation order is not manipulated). In view of this structural feature, transfer models are not able to accurately reproduce learning mechanisms. Using a metaphor, transfer models can be compared to horizontal lines approximating a monotonic function. They could be used to reproduce learning (as horizontal lines could be used to approximate a monotonic function), but they would not perform accurately. Therefore, transfer models are only suitable for replicating participants' performance during the transfer phase exclusively, that is, after a hypothetical learning phase has taken place. Indeed, the transfer phase in an experiment is generally short and because feedback is not provided to participants during this phase, we can assume that learning does not

operate anymore and that participants' performance is (almost) invariant. Conversely, learning models are suitable for reproducing participants' outcomes during both the transfer and learning phases.

On the Present Thesis

The present work is structured on the duality transfer/learning. Since it is easier to work with transfer models than learning models, the description and use of transfer models always precede the description and use of learning models. Among the previously listed categorization models, the present study will only focus on three of them: the GCM (and variants), ALCOVE model, and Component-Cue model (their description and the reason why they were chosen is given in Chapter 3).

1.5 Overview

The present study is based on the hypothesis that investigating the role of presentation order on category learning is relevant to model evaluation. First, it increases our understanding of how learning occurs over time. Since information is sequential, a clear understanding of the temporal aspects underlying category learning sheds light on how the temporal organization of cognitive processes work in general [FTT15; MP15; PM15; SP15; ZM09; ZSP12]. Second, investigating presentation order is important for practical pedagogical applications because modifying the order in which the content is proposed has potentially beneficial consequences for teaching [Dun+13]. Finally, improving our knowledge on how learning takes place when different sequences are presented allows us to suggest strategies for improving the model themselves.

Empirical Approach

There are at least two kinds of strategies that can be adopted in category learning: the similarity-based strategy, by which participants classify new items on the basis of their similarity to the stored exemplars or prototypes of the category, and the rule-based strategy, by which participants classify new items on the basis of rules. These two

strategies have inspired the creation of eponymous within-category similarity- and rule-based presentation orders, with the idea that investigating the similarity- and rule-based orders could help shed light on the nature of the mechanism(s) underlying category learning.

However, the impact of similarity-based versus rule-based presentation order on category learning has been under-explored (see "Similarity-Based vs. Rule-Based Within-Category Orders" in Section 1.3). In addition, the few studies on the topic have been limited to particular contexts. For example, in the past literature, the designs have been restricted to either blocked or random between-category manipulations [EA84; MF09; MF16]. In these limited contexts, rule-based study has been found to be more beneficial than similarity-based study, but it is possible that different stimuli could lead to opposite conclusions.

Therefore, we propose to investigate whether the advantage of a rule-based presentation order can be extended to other contexts. In particular, we focus our attention on examining the impact of similarity- versus rule-based orders when categories are interleaved. Interleaving has proved to be beneficial in a large variety of situations (see "Interleaved vs. Blocked Between-Category Orders" in Section 1.3). However, it is unclear how interleaving interacts with within-category orders, such as similarity- and rule-based orders.

Additionally, following the aspiration to better understand the interaction between different hierarchies of orders (within-category, between-category, etc.), we propose to investigate the impact of a variable versus a constant presentation across blocks. Although the topic of the presentation across blocks has not been addressed in the literature, it has the potential to complement our insights about category learning. On one hand, using constant orders can a priori appear to be the worst idea because such a manipulation has the potential to mislead participants (for instance, focusing participants' attention toward patterns that are inappropriate for the classification), but on the other hand 'presentation order' can become the primary interest and it is possible that constant orders can emphasize order effects.

Another secondary objective is to promote the use of some lesser known statistical tools, such as survival analysis. The majority of the researchers remove from the analysis the participants who did not complete the task or meet the objectives that have been established [CG14b; CG14a; CG14c; CG17; Mea+17; MF16]. Removing unsuccessful participants causes a loss of information which can be detrimental to the domain, especially

if the category structure or presentation order induce difficulties that are detrimental to learning. Therefore, it might be useful to employ more appropriate and more powerful statistical tools to incorporate information provided by the unsuccessful participants.

To recap, we aim to *i)* investigate whether a rule-based study facilitates learning as compared to the similarity-based study in a different variety of category order; *ii)* initiate the investigation regarding the interaction between within-category orders (i.e., stimulus order within categories), between-category orders, and variations of orders across blocks.

Cognitive Modeling Approach

Preliminary findings have shown that similarity- versus rule-based orders (and presentation order in general) can shape the way we perceive categories [MF16], as well as influence our (learning and transfer) performance [EA84; MF09] (see Section 1.3). It is also clear that modeling is the best approach to quantitatively describe the shapes of these representations, as well as the learning process (see "Why Would One Use Cognitive Models?" in Section 1.4; see also [ZH99]).

However, few models are capable of predicting differential performance as a function of presentation order. A recent example is provided by Carvalho and Goldstone's Sequential Attention Theory Model (SAT-M) [CG19]. SAT-M is an exemplar model based on GCM in which the encoding of items depends in part on their temporal proximity. We propose to follow a similar path by providing a new model that integrates an ordinal dimension and by analyzing the temporal dynamics of some relevant categorization models. Our analysis is structured on the basis of the distinction mentioned above: transfer models and learning models.

One intuitive hypothesis is that a model built to account for phenomenon X should be more sensitive to a presentation order that is inspired by phenomenon X. For instance, one could expect that a model implementing similarity processes should increase its sensitivity to similarity when two similar objects are presented contiguously in time. Proximity in time should boost the perceived similarity of the features. Similarly, a model that is supposed to form a general rule on top of exceptions should be more sensitive to a presentation order following that particular structure. But what if the different models cannot fit the data for which they should be more sensitive? Our general hypothesis is that presentation orders can offer a benchmark to evaluate categorization models.

Transfer Models

We develop a new exemplar model based on GCM that accounts for the order in which stimuli are presented. This new model, called Ordinal General Context Model (OGCM), is declined into three versions. The three versions of the OGCM were conceived to investigate whether one of the following orders impacts transfer performance: *i)* the average presentation order received during learning, *ii)* the most frequent presentation order received during learning, and *iii)* the presentation order received during transfer.

These three versions are compared with GCM and GCM-Lag, the latter allowing us to determine whether the integration of a memory-decay mechanism increases the accuracy of the predictions of the transfer performance (details in Chapter 3). The aim of the analysis is twofold: to determine the model that best suits transfer performance and to understand whether the putative differences between within-category types of orders can be captured by the selected models.

Learning Models

The predictions of the learning models evolve over time (see "Ability to Learn" in Section 1.4). This feature should confer learning models the ability to be impacted by presentation order. We propose to test the sensitivity of some learning models to within-category order.

Two learning models are compared: Kruschke's ALCOVE model [Kru92] and Gluck and Bower's Component-Cue model [GB88b]. The selection of these two models is motivated by the fact that they are grounded on distinct psychological mechanisms (a similarity-based mechanism for ALCOVE and a rule-based mechanism for Component-Cue), even though they share a similar mathematical architecture (they both are artificial neural networks). Our aim is to determine which model best suits learning and transfer performance, as well as to investigate their sensitivity toward presentation order.

Chapter outline

The present work is organized on the dual empirical-modeling approach. The empirical approach is developed in Chapter 2, while the modeling approach is developed from Chapter 3 to Chapter 6. The modeling of presentation orders is investigated in

several chapters: Chapter 4 concerns the use of transfer models to reproduce transfer performance, 5 concerns the use of learning models to reproduce learning and transfer performance, and Chapter 6 attempts to use transfer models to reproduce learning performance.

In Chapter 2, we investigate the effect that within-category order (similarity-based versus rule-based) exerts on learning through a series of laboratory experiments. In most cases, learning appears faster in the rule-based order as compared to the similarity-based order.

In Chapter 3 we propose a new exemplar model that integrates the presentation order and present the transfer and learning models that have been compared. Transfer models include the Generalized Context Model (GCM) [Nos86], the new OGCM declined in three versions, and the GCM-Lag [MN95; NKM92; Nos11]. Learning models include the Component-Cue model [GB88b] and the ALCOVE model [Kru92], both declined in two versions (exponential and linear). We also provide the mathematical framework underlying the models and present their likelihood.

In Chapter 4 we develop a methodology for comparing categorization models and we apply it to transfer models. We show that the transfer model that best fits the data is the one that integrates the most frequent presentation order (i.e., the median presentation order that participants received during learning).

In Chapter 5 we apply the inference method developed in Chapter 4 to learning models and we investigate whether models might be more sensitive toward certain within-category orders in particular. We show that individual learning data are best fit by ALCOVE, whereas participants' transfer performance is best fit by Component-Cue (more specifically, Component-Cue provides the best account for two thirds of the participants). Moreover, we show that the generalization patterns of participants who received certain types of orders are captured by different models.

In Chapter 6 we show how to use a transfer model to reproduce learning performance by using either the segmentation or the segmentation/clustering and we apply these two methods to the GCM. We found that there are generally two classes of participants with different learning speeds.

Finally, in Chapter 7 we list the main contribution of the present work, we present some ideas for future directions (in particular ideas for new experiments), and we propose some recommendations.

2

Experimental Data and Preliminary Statistical Analyses

Contents

As seen in Chapter 1 Section 1.3, a variety of studies have investigated the way presentation order influences category learning. While the majority have explored the effect of interleaving versus blocking ([Gol96], [KB08], [KP12], [CG14b]), a more limited number of studies have focused on the impact of within-category presentation order ([EA81; EA84; MF09; MF16]). The totality of the research on within-category order have manipulated the similarity between contiguous examples (for example, maximizing or minimizing the similarity between adjacent examples). Only rarer cases have attempted to explore the rule-based presentation order, a type of order that depends on the logical structure of the to-be-learned categories ([EA84; MF09; MF16]). Moreover, the few research comparing rule-based and similarity-based orders has been limited to specific contexts (e.g., categories were either blocked or randomly alternated).

Goals

In view of the above, the main goal of the present chapter is to further investigate the effects of rule- and similarity-based orders on category learning when additional presentation orders are at play (i.e., between-category orders and across-blocks manipulations). Again, the few studies on the topic showed that the rule-based order was more beneficial than the similarity-based order (in particular contexts). Our aim is to determine whether the advantage of the rule-based order can be extended to other contexts.

Studying the rule- and similarity-based orders serves two purposes. On one hand, it helps better understanding the mechanisms at play during category learning. On the other hand, it helps evaluating categorization models. Again, one plausible hypothesis is that a model integrating a mechanism X should be more sensitive to a presentation order inspired by the mechanism X. Therefore, the manipulation of different presentation orders might allow us to both test the mechanisms underlying the models and determine which model best fits the real world (clarifications will be given in Chapter 4 and 5).

An additional goal is to initiate the investigation on how different hierarchies of orders interact. The majority of studies concerning the way presentation order affects category learning generally focus on a single manipulation of order at a time [CG14b; KP12; KB08; Noh+16]. For example, a study on between-category presentation order would consider different types of between-category orders but use a random presentation order with members within a category. Despite the effectiveness of focusing on a single factor, studying the way different orders interact could shed light on the mechanisms underlying category learning.

Our experimental data investigate how rule- and similarity-based orders influence category learning while manipulating other types of presentation orders (e.g., blocking and across-blocks manipulations), but this data can only be used to partially understand how different (ordinal) contexts influence learning performance. This data does not intend to allow a full factorial comparison between the three hierarchies of orders (i.e., within-category orders, between-category orders, across-blocks manipulations). Instead, it aims to promote further studies investigating the way different types of orders interact by showing attempts to model different types of orders in different contexts.

Outline of this chapter

This chapter provides both the description of the experiments that have been conceived to investigate the rule- and similarity-based orders (Experiment I and II) and the results of a preliminary statistical analysis. The statistical analysis of Experiment I follows the description of Experiment I. The same goes with Experiment II.

2.1 Experiment I

Experiment I dataset was collected by Mathy and Feldman to assess the effects of within-category orders on category transfer (see [MF16]). This dataset has mainly been included in the present thesis as a benchmark for comparing categorization models. Nevertheless, a preliminary statistical analysis of this dataset is provided. Its aim is to enrich the analysis performed by Mathy and Feldman with numerous and more robust statistical tests (details in Subsection 2.1.2).

2.1.1 Data Collection

Participants. The participants were 44 freshman or sophomore students from the University of Franche-Comté (France), who received course credits in exchange for their participation.

Phases. The experiment was composed of two phases: a supervised learning phase (in which participants were trained on the studied categories) and an unsupervised transfer phase (in which participants' ability to generalize their learning was tested).

Categories. Each participant received a single 5-4 category set (Figure 2.1, on the top). The 5-4 category set was first studied by Medin and Schaffer [MS78] and reanalyzed in many subsequent studies ([CN03], [JK05], [JP03],, [LLM07], [Lam00], [MS02], [RH05], [SM00], [Zak+03]). The 5-4 category set is composed of $2^4 = 16$ items placed on a hypercube (more specifically a 4-cube). Only 9 of the 16 items belong to one of the two categories, A and B. The name of this category set is due to the fact that 5 items belong to category A, while 4 items belong to category B. This makes a total of $5 + 4 = 9$ learning

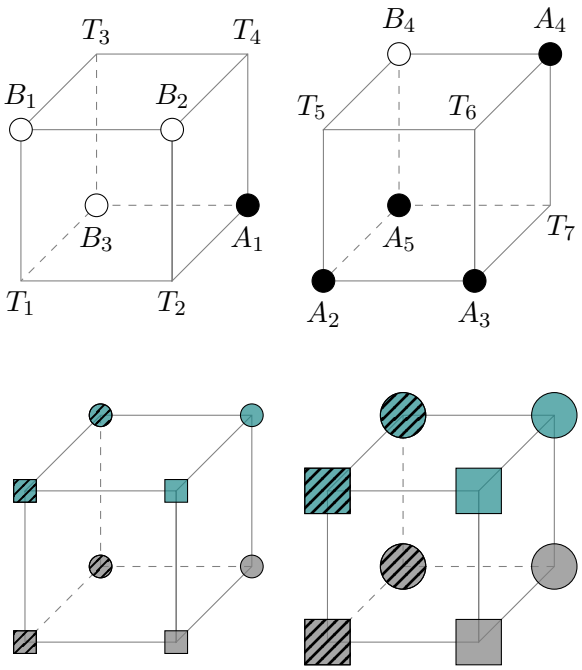


Figure 2.1 – *Illustration of the studied categories and items of Experiment I. At the top, the structure of the 5-4 category set. The examples of category A are indicated by black circles, those of category B by white circles, and transfer item are represented by empty vertices. At the bottom, the illustration of the items of Experiment I. The items varied along four Boolean dimensions (shape, color, size and filling pattern).*

items which are presented in both the learning and transfer phases. The remaining 7 items are transfer items and they are only presented during the transfer phase. In Figure 2.1 (on the top), the examples of category A are indicated by black circles, those of category B are indicated by white circles, while the transfer items are indicated by empty vertices.

Items. Items varied along four dimensions: *shape*, *color*, *size* and *filling pattern*. Each of these dimensions was Boolean, meaning that they could take only two possible values. The two options for each dimension were: square or circle for shape; blue or gray for color; small or big for size; plain or striped for the filling pattern. The combination of these four Boolean dimensions formed $2^4 = 16$ items (Figure 2.1, on the bottom). Each dimension was instantiated by the same physical feature for all participants. Therefore, color differentiated the objects at the top of the hypercube from those at the bottom; shape differentiated the objects at the front from those at the back; size distinguished the

objects in the left cube from those in the right cube; and filling pattern differentiated the right and left objects within the cubes.

Between-category presentation order. During the learning phase, categories were strictly blocked. However, since blocking was not a guarantee for learning (the participant could have pressed the correct keys without looking at the stimuli), blocks in which categories were strictly blocked were alternated with random blocks. In other words, odd blocks (the manipulated blocks) were characterized by a blocked between-category order (i.e., the presentation of category A examples always preceded the presentation of category B examples), while even blocks were characterized by a random between-category order. During the transfer phase, members of category A were randomly alternated with members of category B (random between-category order).

Within-category presentation order. During the learning phase, two within-category presentation orders were used: the rule-based and the similarity-based. For each participant, one of these two presentation orders was randomly chosen beforehand and applied across every odd block of the learning phase. Conversely, in both the even blocks of the learning phase and the transfer phase, members within a category were randomly selected. Among the 44 participants of the experiment, 22 of them were assigned to a rule-based condition, while the remaining 22 were assigned to a similarity-based condition.

In the *rule-based* order, the stimuli were ordered following a "principal rule plus exceptions" structure, meaning that examples obeying the principal rule were presented strictly before the exceptions. The specific "principal rule plus exceptions" structure on which the rule-based order was based is the following: all gray items belong to category A except for the small hatched circle, while all blue items belong to category B except for the big plain circle (see Figure 2.1). Therefore, the main rule was "gray items are members of category A and blue items are members of category B", while the exceptions were the small gray hatched circle and the big blue plain circle. In this "principal rule plus exceptions" structure, items A_1, A_2, A_3, A_5 represented the category A examples obeying to the main rule, while A_4 was the only category A exception. In the same way, B_1, B_2, B_4 were the category B examples obeying to the main rule, while B_3 was the only category B exception. In the rule-based order all the category A items obeying to the dominant rule were presented strictly before the exceptional category A item. The same goes for the category B items. This type of presentation order was thought to

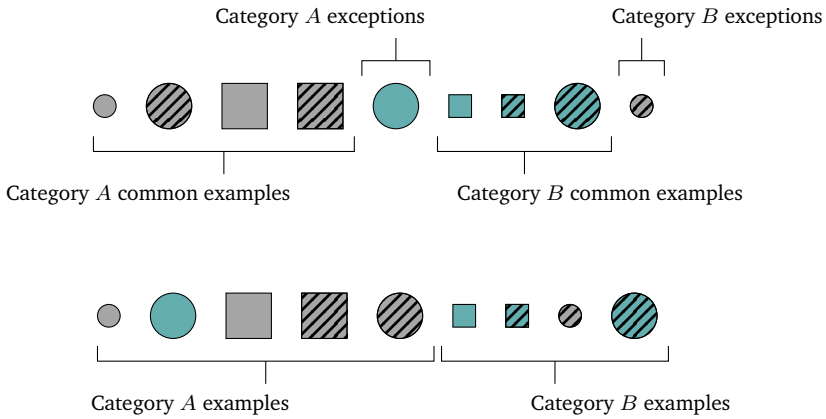


Figure 2.2 – *Examples of blocks following a rule-based (on the top) and a similarity-based (on the bottom) presentation order, using the stimuli and categories of Experiment I. In the rule-based order, the stimuli are ordered following a "principal rule plus exceptions" structure. Conversely, the similarity-based order is designed to maximize the similarity between consecutive stimuli.*

encourage participants to abstract the principal rule shared by all items obeying to it. The items belonging to the principal rule (whether belonging to categories A or B) were randomly selected. An example of block following a rule-based order is given in Figure 2.2 (on the top). Categories are blocked coherently with the between-category order of the experiment.

In contrast, in the *similarity-based* order, members within a category were presented in a way that maximized the similarity between adjacent learning stimuli. The first stimulus was randomly chosen while subsequent stimuli were (randomly) chosen among those that were the most similar to the immediately previous item. Similarity between two items was computed by counting the number of common features that they shared. For instance, the small plain blue circle and the small striped gray square have one single feature in common (small), thus their similarity is 1. Ties were solved randomly. The similarity-based order was thought to have the objective of reinforcing exemplar memorization (see [EA81; EA84]). An example of block following a similarity-based order is given in Figure 2.2 (on the bottom). Categories are blocked coherently with the between-category order of the experiment.

Presentation across blocks. During the learning phase, odd and even blocks were characterized by different across-block presentation orders. In odd blocks, participants received the same sequence of exemplars (constant presentation across blocks), while in

even blocks the sequence of stimuli varied from a block to the next (variable presentation across blocks). During the transfer phase, a variable presentation order across blocks was applied. The rationale here is that testing can only be conducted with random presentations. A summary of all the types of presentation orders manipulated during the learning phase of Experiment I is given in Figure 2.3.

Stop criterion. Participants completed the learning phase when one of the following conditions was satisfied:

i. One sequence of $4 \times (5 + 4)$ consecutive correct responses was given during random blocks, i.e. 4 consecutive correct random blocks (see Figure 2.1 on the top for an illustration of the condition using a smaller number of learning items).

ii. Two distinct sequences of respectively $2 \times (5 + 4) + 1$ and $2 \times (5 + 4)$ consecutive correct responses were given during random blocks (2 consecutive random blocks plus one stimulus and 2 consecutive random blocks). This condition allows the learning phase to end only when no more than two wrong responses in a row were given between the two distinct sequences (see Figure 2.1 on the top for an illustration of the condition using a smaller number of learning items).

When one of the previous conditions have been satisfied, we consider that participants reached the learning criterion. Once participants met the learning criterion, a transfer phase was conducted. The transfer phase was composed of 5 blocks of 16 stimuli (the $5 + 4 = 9$ learning items plus the 7 transfer items).

Procedure and feedback. The categorization task was computer-driven and each participant was individually tested. Participants sat approximately 60 cm from the computer screen and they were briefly instructed before the task began by means of a tutorial. Stimuli were presented one at a time for 1 s on the top half of the computer screen. Category A was depicted as a school bag located at the top right side of the screen and was associated to the up key (to match the visual output). Category B was depicted as a trash can located at the bottom right side of the screen and was associated to the down key.

When stimuli were presented during odd blocks of the learning phase, the correct category label (i.e. “school bag” or “trash”), as long as the corresponding category picture, were displayed for 1 s. The correct category label appeared below the presented stimulus,

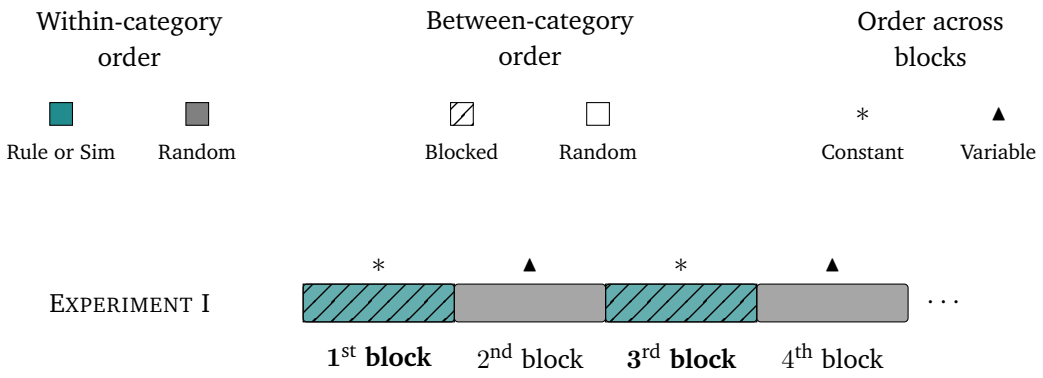


Figure 2.3 – *Illustration of the presentation orders of the learning phase of Experiment I. Color indicates the within-category presentation order: blue for rule- or similarity-based orders, and gray for random order. Filling pattern indicates the between-category presentation order: striped for blocking, and plain for random. Finally, symbols above each block represent the manipulation of stimuli across blocks: a star for a constant presentation and a triangle for a variable presentation. Bold style is used to indicate that the correct classification was given to participants not only after the classification trial but also before it.*

while the corresponding category picture appeared on the right-hand side of the screen. In other words, the wrong category picture disappeared for 1 s, while the right one remained displayed. This instruction was followed by a confirmation phase in which participants had to press the response key corresponding to the right category to make sure the participants was following the training attentively. After the key was pressed, feedback indicating a correct or incorrect classification was displayed for 2 s at the bottom of the screen.

When stimuli were presented during even blocks of the learning phase, participants had to classify it in one of the two categories using the response keys. Once the key pressed, a feedback indicating the correctness of the classification appeared for 2 s at the bottom of the screen. Conversely, during the transfer phase no feedback was provided.

In order to encourage learning, a progress bar representing the participants' score was displayed at the bottom of the screen. The progress bar was composed of $4 \times (5 + 4)$ empty boxes. One point was scored on the progress bar (one empty box was filled) every time participants gave a correct response during the random blocks. On the other hand, the progress bar was reset to zero every time participants gave an incorrect response during the random blocks. An exception to the latter rule was made when participants filled at least half plus one boxes of the progress bar (i.e. $2 \times (5 + 4) + 1$ consecutive correct responses). In this case, every time participants gave an incorrect response (and

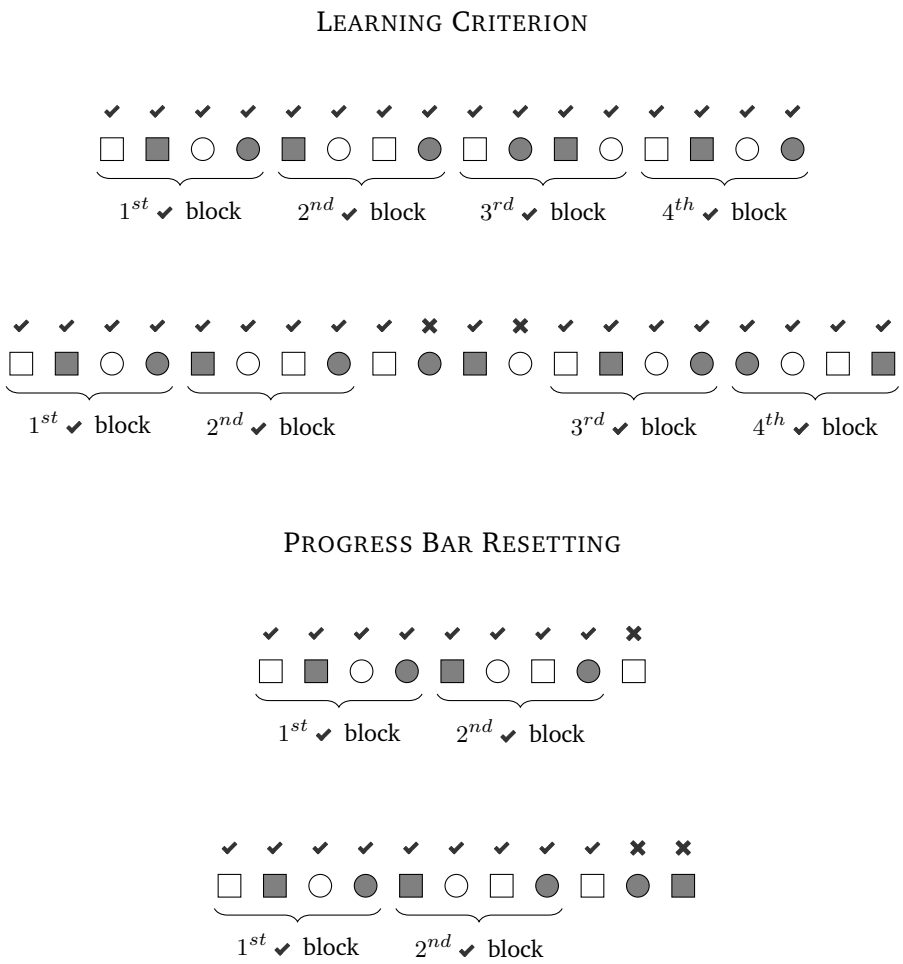


Figure 2.4 – *Illustration of the learning criterion and resetting of the progress bar with fake learning sequences. On the top, illustration of the two conditions (the first on the top, the second on the bottom; see paragraph "Stop criterion") that could allow participants to complete the learning phase. On the bottom, illustration of sequences implying the resetting of the progress bar to zero. The check mark indicates a correct classification of the corresponded stimulus (or block), while the cross mark indicates an incorrect classification of the corresponded stimulus (or block).*

only one consecutive incorrect answer) the progress bar was reset to the half (i.e. the $2 \times (5 + 4)$-th box). The rationale here was to mimic the learning criterion.

Example 2.1. Here, we aim to show the learning criterion defined above using a limited number of stimuli. Let us consider a set of 4 learning items (a white circle, a white square, a gray circle and a gray square) varying along two Boolean dimensions (shape and color). In this case, a block of the learning phase is defined as a sequence of 4 stimuli. Let us transpose the two conditions that could allow participants to complete the learning phase

to this particular case. According to the first condition, the learning phase was completed when participants correctly classified 4×4 consecutive stimuli, i.e. 4 learning blocks (see its visual representation in Figure 2.4, on the top). The second condition is a weaker version of the first one. The learning phase was also completed by correctly classifying two separated sequences of consecutive stimuli of size $4 \times 2 + 1$ and 4×2, respectively. Between the two sequences, participants were not allowed to make two consecutive incorrect classifications (see its visual representation in Figure 2.4, on the top).

In Figure 2.4 (on the bottom) are shown two examples of sequences of responses that imply resetting the progress bar to zero. In the first example, we assume that a participant gave 2×4 consecutive correct responses and then made a mistake. Since the participant did not correctly classify $2 \times 4 + 1$ consecutive stimuli, the progress bar would be reset to zero instead of 2×4. In the second example, a participant gave $2 \times 4 + 1$ consecutive correct responses and then makes two mistakes in a row. The fact that the participant incorrectly classified two stimuli in a row, causes the lost of the earned advantage (2 consecutive correct blocks). ⊠

TO SUM UP *Experiment I: Data Collection*

Participants. There were *44 participants*.

Phases. The experiment was composed of two phases: a supervised learning phase and an unsupervised transfer phase.

Categories. Each participants received a single *5-4 category set*. This category set was composed of 16 items placed on a 4-cube. Nine of the 16 items were learning items (5 belong to category A and 4 belong to category B), while the remaining 7 items were transfer examples.

Items. Items varied along four Boolean dimensions: *shape, color, size* and *filling pattern*. Each dimension was instantiated by the same physical feature for all participants.

Between-category presentation order. The between-category presentation order was only manipulated in the odd blocks of the learning phase. In these blocks categories were blocked, meaning that the presentation of category A examples always preceded that of category B examples.

Within-category presentation order. The within-category presentation order was only manipulated in the odd blocks of the learning phase. Two orders were used: rule-based and similarity-based. In the rule-based order, the stimuli were ordered following a "principal rule plus exceptions" structure, meaning that examples obeying the principal rule were presented strictly before the exceptions. Conversely, the similarity-based order was designed to maximize the similarity between contiguous examples. The within-category order was a between-subject manipulation.

Presentation across blocks. Stimuli across blocks were only manipulated in the odd blocks of the learning phase. In these blocks a constant presentation across blocks was used.

Stop criterion. The learning phase ended when participants reached the learning criterion, meaning they had to correctly classify two consecutive blocks twice without making two or more mistakes in a row between the two sequences. Conversely, the transfer phase was composed of 5 blocks of 16 stimuli.

Procedure and feedback. Each participant was individually tested on a computer-driven task. In the odd blocks of the learning phase, feedback was given before and after the classification trial while in the even blocks feedback was only given after the classification trial. No feedback was provided during the transfer phase.

2.1.2 Analysis of Learning Phase

The analysis conducted by Mathy and Feldman on Experiment I [MF16] showed that rule- and similarity-based orders led to different generalization patterns. Moreover, they detected faster learning in the rule-based condition. Since their study was mainly focused on the transfer phase, the analysis performed on the learning phase was limited to a single statistical test. Therefore, we aim here to reanalyze the learning phase with numerous and more robust statistical tests. This subsection is organized in three parts, each of them tackling the relation between presentation order (rule-based vs. similarity-based) and learning speed from a different angle.

	Successful	Unsuccessful	Total
Rule-based	22	0	22
Similarity-based	21	1	22
Total	43	1	44

Table 2.1 – *Number of successful and unsuccessful participants of Experiment I, depending whether they were assigned to the rule-based or to the similarity-based conditions.*

Unsuccessful participants. Firstly, we investigate whether the number of unsuccessful participants depends on the condition to which subjects were assigned.

Relevant times to estimate learning. Secondly, we focused on a choice of times (that we called relevant times) at which the learning progress can be quantified as a function of within-category order.

Proportion of correct responses. Finally, we investigated whether presentation order influences the evolution of the proportion of correct responses over time.

Unsuccessful Participants

The learning criterion allows us to classify participants into two groups: those who met it and those who did not. The individuals that met the learning criterion are called *successful participants*, while those who did not are called *unsuccessful participants*. The unsuccessful participants usually dropped out the experiment when its duration was close to between 30 minutes and one hour, and when they were not required to stay that long to achieve course credits. With the aim to establish whether the number of unsuccessful participants was related to the within-category presentation order, we ran a Fisher's exact test of independence (see Box 2.1 for a description of the test). In view of the limited number of participants, the Fisher's exact test was preferred to the chi-square test.

Fisher's exact test of independence. Table 2.1 shows the number of successful and unsuccessful participants for both rule-based and similarity-based conditions. We observed that only one participant (following a similarity-based order) did not meet the learning criterion. Although there was only one unsuccessful participant, we decided to run a Fisher's exact test of independence to maintain a coherent structure with Experiment II

(the test was performed using the R function `chisq.test`). As expected, the test was not significant.

Conclusion: The number of unsuccessful participants was not related to the within-category presentation order.

BOX 2.1 *Fisher's Exact Test of Independence*

The Fisher's exact test of independence is a statistical significance test used to determine whether two nominal variables are related. It is usually used instead of the chi-squared test of independence when sample sizes are small. The reason for this is because the significance value of the Fisher's test can be exactly calculated as opposed to that of the chi-squared test that can only be approximated. Therefore, on small samples, the Fisher's test performs better than the chi-squared test.

Let us briefly explain how the test works. Let us suppose that there exist two nominal variables with, respectively, r and s states. The test is based on the contingency table of the two variables, where $O_{i,j}$ ($i \leq r$ and $j \leq s$) are the observed frequencies, $R_i = \sum_j O_{i,j}$ and $C_j = \sum_i O_{i,j}$ are, respectively, the row and column sums, and $N = \sum_i R_i = \sum_j C_j$ are the total sum of the matrix.

The first step consists in declaring the hypothesis. The null hypothesis states that the relative proportions of one variable are independent of the second one, while the alternative hypothesis states the opposite.

The second step is represented by the computation of the conditional probability of getting the actual matrix, given the particular row and column sums, by using the formula:

$$p = \frac{(R_1!R_2!\cdots R_r!)(C_1!C_2!\cdots R_s!)}{N!\prod_{i,j} O_{i,j}!}. \tag{2.1}$$

Equation 2.1 is a multivariate generalization of the hypergeometric probability function.

The third step consists in finding all possible matrices of non-negative integers with the row and column sums R_i and C_j, and calculating the associated conditional probability by means of Equation 2.1. The sum of these probabilities must be equal to 1.

Finally, the last step consists in summing all the probabilities that are equal to or smaller than that of the observed table (which is p). This sum represents the final p-value of the test.

Relevant Times to Estimate Learning

An effective way to compare different speed of learning is to identify a series of relevant times. By relevant times we mean times that vouch for an increase of performance. For instance, the first time (expressed in blocks or trials) at which participants correctly classify two blocks in a row is a relevant time. Each relevant time is associated with the achievement of a particular stage of the learning dynamic. Thus, by comparing specific relevant times of the two conditions (rule-based and similarity-based), we aimed to establish which condition lead to the fastest learning. For each of the selected relevant times (that will be described later), we compared the two following sequences:

$$T_1^{\text{rule}}, \ldots, T_{n_{\text{rule}}}^{\text{rule}} \qquad \text{and} \qquad T_1^{\text{sim}}, \ldots, T_{n_{\text{sim}}}^{\text{sim}}, \tag{2.2}$$

where T_i^{rule} and T_i^{sim} are, respectively, the selected relevant time for the i-th rule-based and the i-th similarity-based participants, while n_{rule} and n_{sim} are, respectively, the number of rule-based and similarity-based participants for whom the selected relevant time is defined. Three analyses were performed to determine whether the two sequences in Equation 2.2 were statistically different: the Wilcoxon-Mann-Whitney test, the Kaplan-Meier estimator, and the Cox model. For each analysis, a set of relevant times was selected and analyzed. These analyses were preferred to more commonly used methods (for instance, the mixed model) because they do not lie on a Gaussian assumption. Indeed, the rate of correct responses in Experiment I ranged from 0 to 1 in 9 steps, which could scarcely be approximated by a normal distribution. Although some versions of mixed models do not assume the distributions to be Gaussian, they generally take long to compute.

Wilcoxon-Mann-Whitney test. The one-sided Wilcoxon-Mann-Whitney test (see Box 2.2 for a description of the test) was run to compare the observed sequences of relevant times. The null hypothesis was the following: the distribution of the relevant times of the rule-based participants is greater than that of the similarity-based participants. The Wilcoxon-Mann-Whitney test was preferred to the Z- and Student's T-tests because *i)*

the number of participants per condition was small (22 participants) and *ii)* no prior knowledge about the underlying distribution was required. The one-sided Wilcoxon-Mann-Whitney test was applied to the following relevant times (the test was performed using the R function `wilcox.test`):

ENDING TIME. The ending time refers to the time at which participants ended the learning phase. For successful participants, it corresponds to the time at which the learning criterion was met, while for unsuccessful participants, it corresponds to the time at which the experiment was dropped. No participants were removed from the study. The p-value of the test was 0.01, showing that the ending time of rule-based participants were significantly smaller than that of similarity-based participants.

LEARNING TIME. The learning time refers to the time at which successful participants met the learning criterion. The only unsuccessful participant was removed from the analysis. The p-value of the test (0.02) indicates a faster learning time for rule-based participants.

FIRST TIME 100%. The first time at which participants correctly classified a block (of 9 stimuli) is another meaningful indicator of participants' learning. Since all participants correctly classified at least one block, nobody was removed from the study. The test was significant with a p-value of 0.01.

FIRST TIME 75%. In the same vein of the previous relevant time, "First time 75%" refers to the time at which participants correctly classify 75% of the stimuli of a block (meaning 6 stimuli over 9). Again, no participants were removed, but this time the test was not significant (the p-value was 0.15).

NEVER UNDER 60% TIME. Finally, we considered the time starting from which participants correctly classify at least 60% of the stimuli of each block (meaning 5 stimuli over 9). No participants were removed and the p-value of the test (0.17) was not significant.

All relevant times were computed in terms of number of trials for best accuracy. Figure 2.5 shows the average of the selected relevant times as a function of the within-category order. To facilitate the comprehension, the average relevant times were here expressed in terms of blocks.

Conclusion: the one-sided Wilcoxon-Mann-Whitney test was significant for Ending time, Learning time and First time 100%, showing faster learning for rule-based participants.

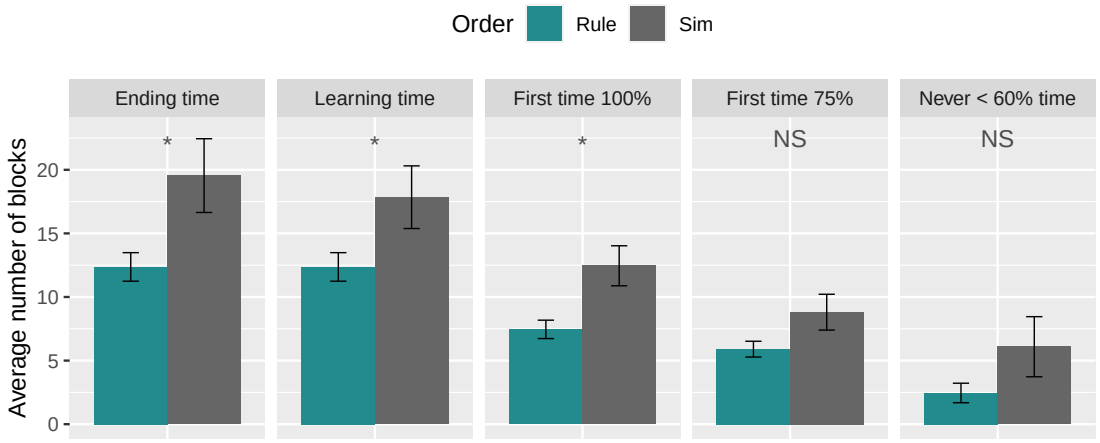


Figure 2.5 – *Average relevant times as a function of the within-category order (similarity-based vs. rule-based). Stars and "ns" symbols indicate the significance of the one-sided Wilcoxon-Mann-Whitney test (see Table 2.2 to map symbols with p-value ranges).*

P-value	0.1 - 1.0	0.05 - 0.1	0.01 - 0.05	0.001 - 0.01	0 - 0.001
Symbol	NS	ns	*	**	***

Table 2.2 – *Mapping from p-value ranges to symbols.*

Box 2.2 *Wilcoxon-Mann-Whitney Test*

The Wilcoxon-Mann-Whitney test (also called Mann-Whitney U test or Wilcoxon rank-sum test) is a non-parametric test that is used to investigate whether two independent samples are likely to derive from populations having the same distribution. The two-sided test is used to determine whether two populations are the same, while the one-sided test is used to detect either a positive or a negative shift (not both at the same time) in one population as compared to the other. The hypothesis of the two-sided test are the following:

H_0 : the distributions of the two populations are equal,

H_1 : the distributions of the two populations are not equal.

In contrast, the null hypothesis of the one-sided test states that there is either a positive or a negative shift (it depends on the direction of the test) in one population as compared to the other. The intuitive idea on which the test is based is the following: if we group the two samples and the elements are uniformly mixed when we reorder them, then the two populations can be considered equal. More specifically, the statistical test is based on the following steps:

i. Merge the two samples into one set and order its elements in ascending order. Assign to each element a rank beginning with 1 for the smallest value. If two or more elements have the same value, assign to them a rank equal to the midpoint of unadjusted rankings. For example the ranks of $(4, 7, 7, 7, 8)$ are $(1, 3, 3, 3, 5)$, while the unadjusted rankings would be $(1, 2, 3, 4, 5)$.

ii. Add up the ranks of the observations coming from the first sample (that we denote by R_1) and those coming from the second sample (that we denote by R_2). Compute the following quantities:

$$U_1 = R_1 - \frac{n_1(n_1 - 1)}{2} \qquad \text{and} \qquad U_2 = R_2 - \frac{n_2(n_2 - 1)}{2},$$

where n_1 and n_2 are, respectively, the number of observations of the first and second samples.

iii. The statistical test U is the smaller value of U_1 and U_2:

$$U = \min\{U_1, U_2\}.$$

The critical value depends on the size of the samples (n_1 and n_2) and on the level of significance (generally $\alpha = 0.05$). If the observed statistic U is smaller than the critical value, then the null hypothesis is rejected.

Kaplan–Meier survival curves and Log-Rank test. There is a less common branch of statistics, called survival analysis, that we thought could best serve our purpose. Survival analysis is a set of statistical techniques used to investigate the expected duration of time until an event of interest occurs (see Box 2.3 for further details). In our case, the event of interest is the attainment of the learning criterion, which marks the successful completion of the learning phase. However, for some individuals, the event may not be observed within the time period of the study (i.e., when participants are unsuccessful). Survival

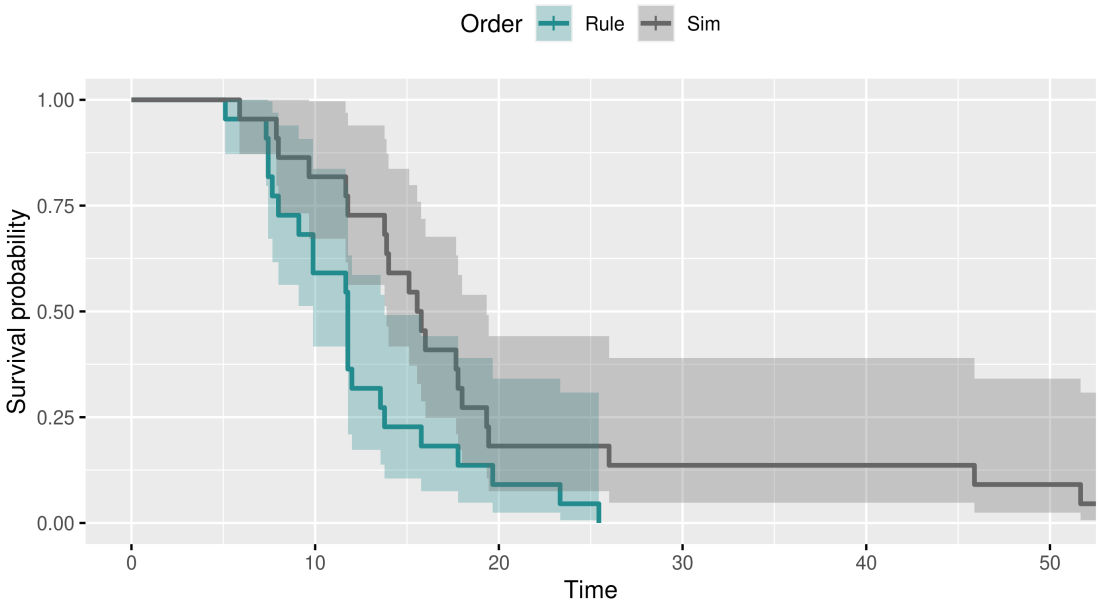


Figure 2.6 – *Illustration of the Kaplan-Meier survival curves of Experiment I as a function of the within-category order. Survival curves represent the probability that participants have not yet reached the learning criterion at a certain time. The transparent areas around the survival curves represent the 95% confidence intervals. The log-rank test showed a significant difference between the two survival curves (p-value=0.02).*

analysis is able to take into account individuals for which the event did not occur, also called the censored observations.

Although the use of survival analysis was not required in Experiment I (a single participant was unsuccessful), we performed it anyway to maintain a common outline between Experiment I and II. However, we anticipate that the use of survival techniques will add a considerable value to the analysis of Experiment II (in which the number of unsuccessful participants was high).

A quantity of interest in survival analysis is the survival probability, also called survival function or survival curve. The survival probability indicates, at a certain time t, the probability that a subject survives longer than time t. In our case, it represents the probability that a participant has not yet reached the learning criterion at time t. One of the most common method to estimate the survival function is the Kaplan-Meier estimator [KM58] (see Box 2.4 for further details).

Figure 2.6 shows the Kaplan-Meier survival curves as a function of the within-category order (survival curves were computed using the R function `survfit`). The distance between the survival curves of the two conditions (similarity-based vs. rule-based) can be quantified using a log-rank test (see Box 2.5 for further details). The log-rank test performed on the similarity- and rule-based survival curves was significant (p-value=0.02; the test was performed using the R function `survdiff`).

Conclusion: the analysis of the Kaplan-Meier survival curves showed a relation between learning speed and within-category presentation order, with rule-based participants having the highest probability to reach the learning criterion.

Box 2.3 *Survival Analysis*

Survival analysis is a branch of statistic which is used to investigate the time that it takes for an event of interest to occur, such as failure in machines or death in biological organisms. However, in some cases the event may not be observed, producing the so-called censored observations. Survival analysis can account for these censored observations.

Let us suppose that the event of interest involves individuals (n individuals to be more specific). The times that we observe are the result of the interaction of two samples which we do not have access to. These samples represent, respectively, the time at which the event of interest occurs, and the time at which the censor occurs. We denote these samples by

Y_s is the time at which the event of interest occurs for the s-th individual,
C_s is the time at which the censor occurs for the s-th individual,

where $s = 1, \ldots, n$. The samples to which we have access are the sample of the observed times (denoted by T_s) and the sample of the individuals that are censored (denoted by δ_s). We define these two observable samples as follows ($s = 1, \ldots, n$):

$$T_s = \min\{Y_s, C_s\} \qquad \text{and} \qquad \delta_s = \begin{cases} 1 & \text{if } Y_s \leq C_s \\ 0 & \text{otherwise} \end{cases}$$

The survival analysis uses these two observable samples to estimate two related probabilities: the survival function (also called survival curves), which is the probability that an individual survives longer than a certain time; and the hazard probability (also called hazard curves), which is the probability that the event occurs at a certain time. The widespread methods used in survival analysis are:

i. The Kaplan-Meier plots to visualize survival curves.

ii. The log-rank test to compare the survival curves of two groups.

iii. The Cox proportional-hazards model regression to asses the influence of one or multiple variables on hazard probability.

Box 2.4 *Kaplan-Meier Estimator*

The Kaplan-Meier method [KM58] is a non-parametric method used to estimate the survival probability from observed survival times. The estimator of the survival probability at time t_i is computed as follows:

$$S(t_i) = S(t_{t-1}) \left(1 - \frac{d_i}{n_i}\right),$$

where $S(t_{i-1})$ is the probability of being alive at time t_{i-1}, n_i is the number of individuals that are alive just before t_i, and d_i is the number of events at time t_i. The time is initialized at 0 ($t_0 = 0$) and the survival probability at 1 ($S(0) = 1$). The estimated survival probability $S(t)$ is a step function that changes value only at the time of each event. The plot of the Kaplan-Meier survival curves as a function of time provides a summary of the data that can be used to estimate measures such as median survival time. It can also be useful to compare in a qualitative way the survival probability of two or more groups.

Box 2.5 *Log-Rank Test*

The log-rank test is the most widely used method of comparing survival curves. It is a non-parametric test, which means that it makes no assumptions about the

survival distributions. The null hypothesis is that there is no difference in survival between the two groups, while the alternative hypothesis claims that survival curves are not identical. The log rank test compares the observed number of events in each group to what would be expected if the null hypothesis were true (i.e., if the survival curves were identical). The log rank statistic is approximately distributed as a chi-square statistical test.

Cox proportional-hazards model. The Cox model [Cox72] is another survival analysis technique (see Box 2.6 for further details). Again, although Experiment I does not require the use of survival techniques, we performed it for purposes of coherency. A key concept in the Cox model is the hazard probability, also called hazard function. The hazard function indicates the risk that the event of interest occurs at a certain time. In our case, it represents the probability that participants meet the learning criterion at a specific time.

Figure 2.7 shows the result of the application of the Cox proportional-hazards model on Experiment I (Cox anlysis was performed using the R functions `coxph`). The rule-based condition was the reference condition and, consequently, had a hazard ratio of 1. The

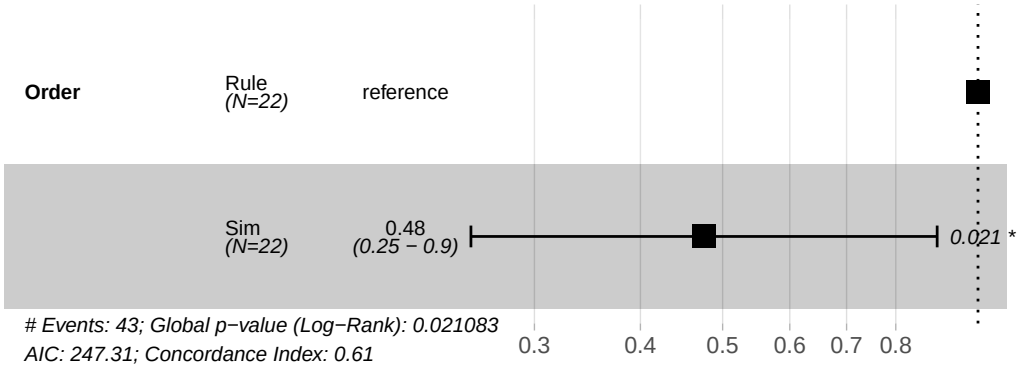


Figure 2.7 – *Results of the Cox proportional-hazards model on Experiment I. The rule-based condition is the reference condition and, consequently, its hazard ratio is equal to 1. Conversely, the similarity-based condition is the opposite condition and its hazard ratio is equal to 0.48 (it is displayed below the word "reference"). The numbers within the brackets just below the hazard ratio represent the 95% confidence interval. The number on the right side of the graph is the p-values of the Wald test assessing the significance of the model.*

similarity-based was the alternative condition and had a hazard ratio of 0.48, meaning that similarity-based participants had lower chances to reach the learning criterion as compared to rule-based participants. More precisely, the similarity-based condition reduces the hazard by a factor of .48 (or 52%).

Another meaningful indicator was the statistical significance of the model, given by three tests (the likelihood-ratio test, the Wald test and the log-rank test). All three tests were significant (p-value=0.02), meaning that the probability to reach the learning criterion was strongly related with the within-category order.

Conclusion: the Cox analysis showed that rule-based participants had higher probability to meet the learning criterion as compared to similarity-based participants.

Box 2.6 *Cox Proportional-Hazards Model*

The Cox model [Cox72] is a survival analysis technique used to examine the effect of several factors on survival probability. In other words, this method allows us to examine how some specified factors influence the rate at which the event of interest happens at a certain time. In the Cox model the hazard function $h(t)$ (which is the probability that the event of interest occurs at a certain time) is expressed as a function of one or more variables $x_1, \ldots, x_n$ called covariates or factors:

$$h(t) = h_0(t)e^{\beta_1 x_1 + \ldots + \beta_n x_n},$$

where $n \in \mathbb{N}$ is the number of factors, $\beta_1, \ldots, \beta_n$ are the coefficients of the variables (they measure the impact of covariates on hazard), and h_0 is the baseline hazard. The Cox model can also be written as a multiple linear regression of the logarithm of the hazard function on the variables x_i.

The quantities e^{β_i} are called hazard ratios. A hazard ratio greater than 1 (or, equivalently, a value of β_i greater than 0) indicates that as the value of the i-th covariate increases, the event hazard increases and thus the length of survival decreases. On the contrary, a hazard ratio smaller than 1 (or, equivalently, a value of β_i smaller than 0) indicates that as the value of the i-th covariate increases, the event hazard decreases and thus the length of survival increases.

The Cox model is based on the assumption according to which the hazard curves of groups of individuals should be proportional and cannot cross. In other words, if an individual has a risk of death at some initial time that is twice as high as compared to another individual, then at later times the risk of death remains twice as high. This assumption should be tested before the application of the model.

Proportion of Correct Responses

The aim here is to compare (at each block) the proportion of correct responses scored by rule- and similarity-based participants during the learning phase. An issue we had to confront was the different duration of the learning phase among participants (see Box 2.7 to understand the issue of dealing with learning phases of different lengths). A first solution could have been to complete the data with 100% of correct responses. However, the unsuccessful participants would have made this solution inadequate. A second solution could have been to remove the unsuccessful participants and complete the data of the remaining subjects with 100% of correct responses. However, the implementation of this solution would have introduced a considerable bias (especially in Experiment II, where the number of unsuccessful participants is higher than Experiment I).

Therefore, we adopted the following alternative solution: we considered the earliest time at which unsuccessful participants dropped the experiment and complete the data until

	Successful	Unsuccessful	All together
Fastest	5	**55**	5
Average	15	55	15
Median	13	55	13
Slowest	51	55	55

Table 2.3 – *Number of blocks that the fastest, the average, the median and the slowest participants of Experiment I took to end the learning phase, depending on the class of individuals they belong to (successful, unsuccessful, or all together). By ending the learning phase we mean reach the learning criterion, in the case of successful individuals, or drop out the experiment, in the case of unsuccessful participants. Bold letters are used to indicate the limit time. The numbers were rounded to their nearest smallest integer.*

this time with 100% of correct responses. The time at which the fastest unsuccessful participant dropped the experiment is called *limit time* and denoted by L.

Avoiding the removal of unsuccessful participants represents a first advantage of this solution. A second advantage is represented by the fact that the completion of data concerns only successful participants (and it is reasonable to think that successful participants will continue to correctly classify stimuli even after reaching the learning criterion). However, a limitation of this method is that only the data preceding the limit time are analyzed.

Table 2.3 shows the number of blocks that the faster, the slowest, the median and the average participant took to end the learning phase. The fastest unsuccessful participant ended the learning phase after the slowest successful participant (55 vs. 51 blocks), allowing us to consider the entire learning dynamics. After addressing the issue of the variable length of the participants' learning phase, the Wilcoxon-Mann-Whitney test was run to asses the influence of the order conditions on the proportion of correct responses.

Box 2.7 *Variable Length Issue*

Let us suppose that we want to compare the performance of two groups of people under two conditions, I and II, on a particular task. Let us assume that the task ends when a participant correctly responses all the questions. To compare the two groups, at time t_1 and t_2 (with $t_1 < t_2$), we evaluate the participants' outcome by giving:

0 if the participant correctly responses less than 50% of the questions.

1 if his percentage of correct responses is between 50% and 99%.

2 if the participant correctly responses 100% of the questions.

Let us assume that, at time t_1, all participants under condition I scored 2, except for one individual who scored 0, while all participants under condition II scored 1. Therefore, only one single participant under condition I was evaluated at time t_2 (the other participants ended the task before the time t_2). Let us suppose that, at time t_2, the only participant under condition I scored 0, while all participants under condition II scored 1. Thus, if we do not take into consideration the participants who ended the task, we could conclude that, at time t_2, the participants under condition I performed better than those under condition II (which is not the case).

This argument shows that analyzing data in which participants take different times to end the task could lead to the wrong conclusion.

Wilcoxon-Mann-Whitney tests with Benjamini-Hochberg correction. Again, the use of the Wilcoxon-Mann-Whitney test is motivated by the absence of prior knowledge about the distribution of the correct responses. The application of the Wilcoxon-Mann-Whitney test aimed to compare the following two sequences:

$$R_1^{\text{rule},j}, \ldots, R_{n_{\text{rule}}}^{\text{rule},j} \qquad \text{and} \qquad R_1^{\text{sim},j}, \ldots, R_{n_{\text{sim}}}^{\text{sim},j},$$

where $R_i^{\text{rule},j}$ and $R_i^{\text{sim},j}$ are, respectively, the proportion of correct responses given by the i-th rule-based and similarity-based participants at the j-th block ($1 \leq j \leq L$), and n_{rule} and n_{sim} are, respectively, the number of rule-based and similarity-based participants. The null hypothesis assumed that similarity-based participants had a higher proportion of correct responses as compared to rule-based ones.

The one-sided Wilcoxon-Mann-Whitney test was applied exclusively to the even blocks of the experiment. This choice is motivated by the fact that in odd blocks feedback were provided before the classification trials.

Since multiple comparisons were at play, a correction procedure was mandatory. Therefore, the Benjamini-Hochberg procedure [BH95] was applied to the results of the Wilcoxon-Mann-Whitney tests (see Box 2.8 for the description of the Benjamini-Hochberg procedure).

Figure 2.8 (on the top) shows the p-values of the Wilcoxon-Mann-Whitney test performed on the even blocks of Experiment I until the limit time. The p-values are ordered from the smallest to the largest. Figure 2.8 (on the top) also shows the straight lines associated with the Benjamini-Hochberg procedure at a significance level of 0.05 and 0.11. If a significance level of 0.05 is considered, all tests are accepted (there are no points under the black line). Conversely, if a significance level of 0.11 is considered, 20 tests over 27 are rejected (the 20-th point appears below the gray line). However, 0.11 is a too high significance level, which led us to the conclusion that no difference between similarity- and rule-based learning curves was found. Figure 2.8 (on the bottom) shows the average proportion of correct responses as a function of the within-category order. Asterisk symbols represent the rejected blocks with a significance level of 0.11.

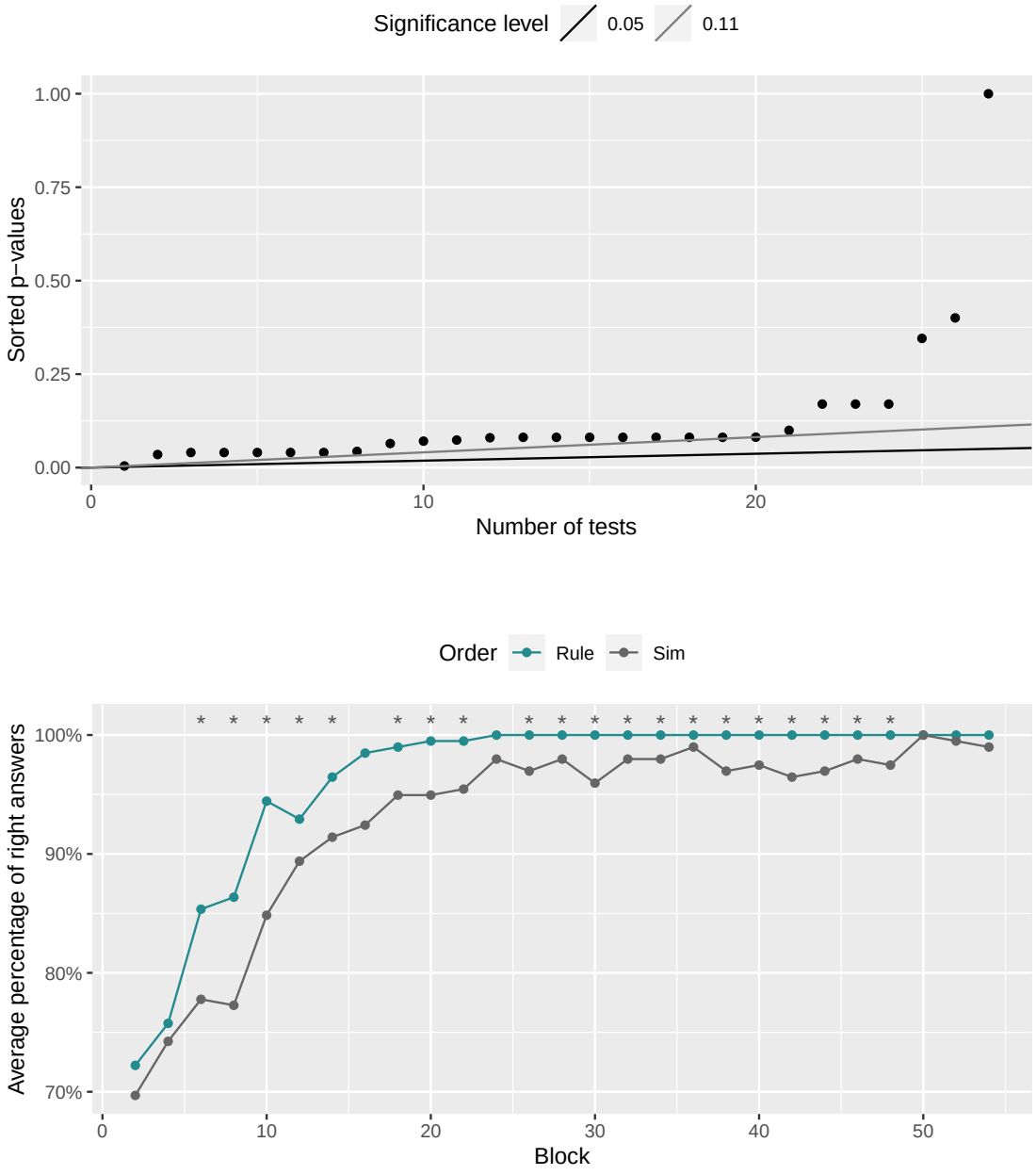


Figure 2.8 – *Result of Wilcoxon-Mann-Whitney tests with Benjamini-Hochberg correction. On the top: p-values of the Wilcoxon-Mann-Whitney test performed to compare the proportion of correct responses between rule-based and similarity-based participants of Experiment I. The test was performed on each block until the limit time (which is the block at which the fastest unsuccessful participant dropped out the experiment). The p-values are ordered from the smallest to the largest. The straight lines are associated to the Benjamini-Hochberg procedure with a significance level of, respectively, 0.05 and 0.11. On the bottom: average proportion of correctly classified items for both rule-based and similarity-based participants of Experiment I. The asterisk symbol denotes the blocks that have been rejected by the test with a significance level of 0.11.*

Conclusion: Although the average learning curve of rule-based participants is always above the average learning curve of similarity-based participants, the Wilcoxon-Mann-Whitney tests with a Benjamini-Hochberg correction found no significant difference between the curves.

BOX 2.8 *Benjamini-Hochberg*

The Benjamini-Hochberg procedure [BH95] is a statistical tool for controlling the false discovery rate (the chances to incorrectly reject the test) in multiple testing experiments. Before tackling the description of this technique, it comes natural to ask ourselves: why do we need to control the false discovery rate when we run a test multiple times? The answer is that the chance of erroneous rejections increases when we perform multiple comparisons. Let us take an example to illustrate this phenomenon. When a single test is performed at the 5% level, there are 5% chance of incorrectly rejecting the null hypothesis. However, if 20 independent tests are conducted, the probability to not have erroneous rejections is $0.95^{20} = 0.36$. This means that there are 64% chance to incorrectly reject the null hypothesis at least one time! This shows that it is of vital importance to control the false discovery rate. Let us say that we performed n tests and that we want to apply the Benjamini-Hochberg procedure in order to limit the false discovery rate at $\alpha = 5\%$. This procedure is based on the following steps:

i. Order the p-values of the tests in ascending order: $p_{(1)} \leq \ldots \leq p_{(n)}$.

ii. Assign a rank to each p-value, beginning from 1 for the smallest value until n for the largest one.

iii. Find the p-values $p_{(i)}$ that satisfy the following condition: $p_{(i)} \leq \frac{\alpha i}{n}$. In a plot with the rank on the x-axis and the p-value on the y-axis, this step corresponds to detect the points which are below the straight line of equation $y = \frac{\alpha}{n}x$.

iv. Set k the largest p-value satisfying the previous condition:

$$k = \max_{1 \leq i \leq n} \left\{ p_{(i)} \leq \frac{\alpha i}{n} \right\}.$$

v. Reject the tests whose rank is smaller than or equal to k, $p_{(1)} \leq \ldots \leq p_{(k)}$.

TO SUM UP *Experiment I: Analysis of Learning Phase*

Experiment I dataset was already used by Mathy and Feldman [MF16] to investigate the influence of within-category order (rule-based vs. similarity-based) on generalization patterns. The aim was to reanalyze the learning phase of Experiment I to investigate the effects of rule- and similarity-based orders on category learning using numerous and more robust statistical tests. The analysis was organized in three parts: *i)* determining whether the number of unsuccessful participants was related to the within-category order, *ii)* analyzing a set of relevant times to compare the learning speed of rule- and similarity-based participants, and *iii)* comparing the evolution of correct responses of rule- and similarity-based participants.

Unsuccessful Participants

Fisher's exact test of independence (not significant)**.** A Fisher's exact test of independence was performed to determine whether within-category order and number of unsuccessful participants were related. The test was not significant.

Relevant Times to Estimate Learning

Both classic methods and survival analysis techniques were used to analyze a set of relevant times.

Wilcoxon-Mann-Whitney test (significant overall)**.** Five relevant times were considered: the Ending time (i.e., the time at which participants ended the learning phase), the Learning time (i.e., the time at which successful participants met the learning criterion), the First time 100% (i.e., the first time at which participants correctly classify a block), the First time 75% (i.e., the first time at which participants correctly classify 75% of the items within a block), and the Never time 60% (i.e., the time from which participants correctly classify at least 60% of the items within each block). A Wilcoxon-Mann-Whitney test was performed to compare the relevant times of rule- and similarity-based participants. The test was overall significant, showing faster learning in the rule-based order.

Kaplan-Meier survival curves and Log-Rank test (significant)**.** We estimated the survival curves of both rule-based and similarity-based participants using the Kaplan-Meier estimator. The distance between the two survival curves was then assessed using a log-rank test. The test showed that participants in the similarity-based order were less likely to reach the learning criterion as compared to participants in the rule-based order.

Cox proportional-hazards model (significant)**.** The Cox model was used to express the hazard probability as a function of the within-category order. The model was significant and confirmed the result of the previous survival analysis, meaning that participants in the rule-based order had higher probability to reach the learning criterion as compared to participants in the similarity-based order.

Proportion of Correct Responses

The dataset was analyzed until the limit time (the block at which the fastest unsuccessful participant dropped the experiment) and completed with 100% of correct responses. This allowed a re-normalization of the duration of participants' learning phase.

Wilcoxon-Mann-Whitney tests with Benjamini-Hochberg correction (not significant)**.** A Wilcoxon-Mann-Whitney test was performed to compare the proportion of correct responses of rule- and similarity-based participants at each block. The result of the test was then corrected using the Benjamini-Hochberg procedure. The smallest significance level that allowed the rejection of at least one test was equal to 0.11 (20 rejected tests over 27, 74% of the tests), which was too high to be significant.

2.1.3 Analysis of Transfer Phase

A large panel of tests has already been performed by Mathy and Feldman on the transfer phase of Experiment I (see [MF16]). They provided evidence supporting the influence of within-category presentation order on generalization patterns. Our aim is to enrich their panel of tests with an additional analysis.

Principal component analysis with Wilcoxon-Mann-Whitney test. Since generalization patterns were investigated, the analysis was focused on transfer items exclusively. The first step of the analysis consisted in computing the proportion of time that each participant classified each transfer item into category A. This operation leaded to the creation of a table in which each row corresponded to a participant and each column corresponded to a transfer item. The entry (i, j) of the table contained the proportion of times that participant i classified the transfer item j into category A during the transfer phase.

Because of the high number of columns (there were 7 transfer items and thus 7 columns), detecting some patterns that distinguished the rule- from the similarity-based participants would have been inaccessible. A viable solution was to reduce the dimension of the space (the number of columns) in order to visualize the data.

The principal component analysis provided a solution for reducing the dimension of the space while preserving the highest quantity of information (see Box 2.9 for an introduction to the principal component analysis). This technique allowed us to determine the directions that account for the highest variability of the data and project the data along them.

Figure 2.9 shows the result of the principal component analysis on the first and second components as a function of the within-category order. On the first component, the majority of the similarity-based participants are located on the left side of the plot, while the majority of the rule-based participants are located on the right side of the plot. The one-sided Wilcoxon-Mann-Whitney test was performed to establish whether the difference in location was statistically significant. The test was significant (p-value=0.02), showing that the generalization patterns of rule-based participants were different from those of similarity-based participants.

Moreover, by looking at how the first component was expressed as a function of the transfer items, we were able to further interpret the test. The first component was expressed as follows:

$$\begin{aligned}\text{Comp.1} = {} & 0.3 \cdot p_A(T_1) + 0.2 \cdot p_A(T_2) - 0.3 \cdot p_A(T_3) \\ & - 0.5 \cdot p_A(T_4) - 0.5 \cdot p_A(T_5) - 0.5 \cdot p_A(T_6) - 0.02 \cdot p_A(T_7), \quad (2.3)\end{aligned}$$

where $p_A(T_i)$ represents the proportion of times that participants classified transfer item T_i into category A.

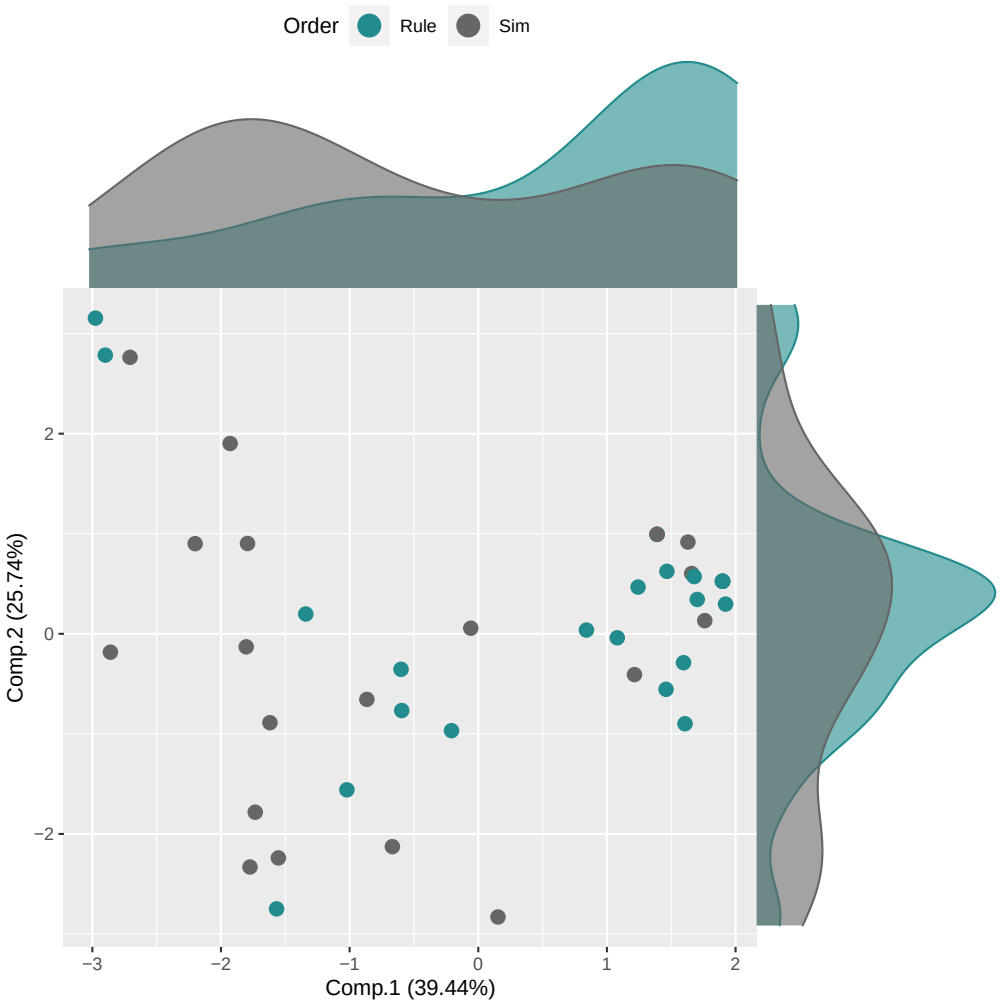


Figure 2.9 – *Projection on the first and second components of the proportion of times that participants classified each one of the transfer item in category A, depending on the condition to which participants were assigned. On the top side and on the right side of the graph, we plot the density of the points on, respectively, the first and the second components.*

The observation that the term involving $p_A(T_7)$ was negligible (its impact was 10 times smaller than the other terms) led us to remove it. The fact that the location of rule-based participants was shifted to the right as compared to the location of similarity-based participants meant that either the positive terms of Equation 2.3 were higher for rule-based participants, or the negative terms of Equation 2.3 were smaller for rule-based participants (or both).

The interpretation in terms of classification probability was the following: either rule-based participants classified items T_1 and T_2 into category A more frequently than

similarity-based participants, or rule-based participants classified items T_3, T_4, T_5 and T_6 into category A less frequently than similarity-based participants (or both).

This configuration is consistent with a rule-based retrieval in which items T_1 and T_2 are classified into category A and items T_3, T_4, T_5 and T_6 are classified into category B. This result showed that the generalization patterns of rule-based participants was closer to rule-based retrieval than the generalization patterns of similarity-based participants.

Conclusion: the location of rule-based participants on the first and second components of the PCA (that was applied to the proportion of times that participants classified each one of the transfer items into category A) was significantly shifted to the right as compared to the location of similarity-based participants. Moreover, the generalization patterns of rule-based participants was closer to rule-based retrieval than the generalization patterns of similarity-based participants.

BOX 2.9 *Principal Component Analysis*

Principal component analysis (PCA) is a statistical technique used to highlight strong patterns in a dataset. If a dataset is composed of more than three variables, then it could be very difficult to visualize the multi-dimensional space in which the observations are embedded. Yet, the visualization of the data could be an useful tool for patterns finding.

The principal component analysis allows us to determine the directions that account for the highest variability of the data and to project the observations along them. In other words, this technique expresses the data in terms of new variables (called principal components) on which the observations are the most spread out. These new variables correspond to a linear combination of the originals. The number of principal components is less than or equal to the number of original variables. In most cases, the first three principal components are able to explain a very high percentage of the data. This allows us to easily visualize the observations with minimal loss of information.

Mathematically speaking, the PCA is an orthogonal linear transformation that transforms the data to a new coordinate system such that the first coordinate accounts for the greatest variance of the data, the second coordinate for the second greatest

variance of the data, and so on. If we denote by X the data matrix (of dimension $n \times m$) in which each row x_i represents an observation, then the transformation is defined by a $m \times p$ matrix W that maps each row vector x_i into a new vector z_i:

$$Z = X \cdot W,$$

where Z is the $n \times p$ matrix with the new coordinates and the columns of the matrix W (the principal components) are eigenvectors of the matrix $X^T \cdot X$. The eigenvalues of the matrix $X^T \cdot X$ measure the amount of variation retained by each principal component. It is also important to point out that the dataset is generally scaled before the application of the technique (variables are scaled to have mean zero and standard deviation one).

2.2 Experiment II

Experiment II was designed to investigate the effects of rule- and similarity-based orders in three different contexts: when categories are randomly alternated and a variable across-blocks presentation is considered (Random-Variable); when categories are randomly alternated and a constant across-blocks presentation is considered (Random-Constant); and when categories are blocked and a constant across-blocks presentation is considered (Blocked-Constant).

Although these three different contexts (i.e., Random-Variable, Random-Constant, and Blocked-Constant) were thought to be three independent experiments, we considered them as (ordinal) contexts of a same experiment (Experiment II). This choice was motivated by practical reasons (a unique statistical analysis was more convenient) as well as the wish to (partially) explore how different contexts influenced the learning speed.

Again, our goal here is to compare rule-based vs. similarity-based within-category orders when other types of orders (across-blocks manipulations and between-category orders) are at play. Additionally, a limited exploration of the effects of multiple types of orders (within-category orders, between-category orders, and manipulations across blocks) on category learning is initiated. To recap, Experiment II was characterized by the three following conditions:

RANDOM-VARIABLE. This condition investigated the effect of within-category orders (rule-based vs. similarity-based) when categories were randomly alternated and a variable across-blocks presentation was considered. We often refer to this condition as R-V.

RANDOM-CONSTANT. This condition investigated the effect of within-category orders (rule-based vs. similarity-based) when categories were randomly alternated and a constant across-blocks presentation was considered. We often refer to this condition as R-C.

BLOCKED-CONSTANT. This condition investigated the effect of within-category orders (rule-based vs. similarity-based) when categories were blocked and a constant across-blocks presentation was considered. We often refer to this condition as B-C.

2.2.1 Data Collection

Participants. The participants were 68, 22, and 46 freshmen or sophomores of the University of Franche-Comté (France) for respectively the R-V, the R-C, and the B-C contexts. All students received course credits in exchange for their participation. The data was collected by F. Mathy.

Phases. The experiment was composed of a single supervised learning phase.

Categories. According to Feldman's classification [Fel03], participants were tested on a single $12_{4[8]}$ concept. This concept is composed of $2^4 = 16$ items placed on an hypercube (more specifically a 4-cube). According to the concept, 8 of the 16 items belong to category A and the remaining 8 items belong to category B. All 16 items are learning items. The concept $12_{4[8]}$ is illustrated in Figure 2.10 (on the top), where the category A examples are indicated by black circles, while the category B examples are indicated by white circles. In the $12_{4[8]}$ notation, the $[8]$ stands for the number of items belonging to category A (8 category A items), the 4 stands for the dimension of the concept (a 4-cube), and the 12 is an arbitrary label identifying this concept among all the $4[8]$ concepts (4-dimensional concepts with 8 items belonging to category A).

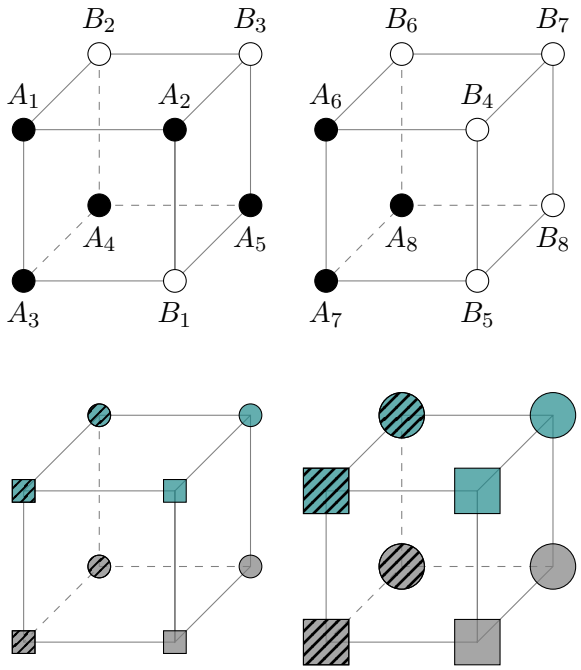


Figure 2.10 – *Illustration of the categories and items of Experiment II. On the top, the illustration of the* $12_{4[8]}$ *concept (according to Feldman's classification [Fel03]). Category* A *items are indicated with black circles, while category* B *items are indicated with white circles. The notation* $12_{4[8]}$ *refers to the fact that this concept is the* 12^{th} *in the Feldman's list of* 4*-dimensional concepts consisting of* 8 *items belonging to category* A*. On the bottom, an example of 16 items presented to participants in Experiment II. The items varied along four Boolean dimensions (shape, color, size and filling pattern).*

Items. Items varied along four dimensions: *shape*, *color*, *size* and *filling pattern*. Each of these dimensions was Boolean, meaning that only two values were available. The choice of the two values for each dimension was chosen at random among these options: triangle, square or circle for shape; blue, pink, red or green for color; small or big for size; plain and striped for the filling pattern. The combination of these four dimensions formed $2^4 = 16$ items (Figure 2.10, on the bottom). Each dimension was not instantiated by the same physical feature. For instance, color could differentiate the items at the top of the hypercube from those at the bottom for a specific participant, while it could differentiate the objects at the front of the hypercube from those at the back for another participant.

Between-category presentation order. In both the Random-Variable and Random-Constant contexts, categories were randomly alternated. Conversely, in the Blocked-

Constant context categories were strictly blocked, meaning that that the presentation of category A items always preceded the presentation of category B items. Moreover, since blocking categories did not guarantee learning, blocks where categories where strictly blocked were alternated with random blocks.

Within-category presentation order. As in Experiment I, a rule-based and a similarity-based orders were used. The within-category order was a between-subject manipulation. In both the Random-Variable and Random-Constant contexts, one of the two within-category orders was randomly selected and applied across every block. Conversely, in the Blocked-Constant context, one of the two within-category orders was randomly selected and applied across every odd block. In each context, half of the participants were assigned to the rule-based order and half of the participants to the similarity-based order.

Again, in the *rule-based* order, stimuli were presented following a "principal rule plus exceptions" structure, meaning that examples obeying the principal rule were presented strictly before the exceptions. The "principal rule plus exceptions" structure of Experiment II was the following: all striped items belong to category A except for the blue circles, while all plain items belong to category B except for the small blue square and the small gray circle (see Figure 2.10, on the bottom). Therefore, the principal rule is "the striped items are members of category A and the plain items are members of category B", while the exceptions are the small striped blue circle, the big striped blue circle, the small plain blue square and the small plain gray circle.

In this "principal rule plus exceptions" structure, items A_1, A_3, A_4, A_6, A_7, A_8 are the category A examples obeying to the principal rule, while A_2 and A_4 are the category A exceptions. In the same way, B_1, B_3, B_4, B_5, B_7, B_8 are the category B examples obeying to the principal rule, while B_2 and B_6 are the category B exceptions. Both the members obeying the principal rule and the category B exceptions were presented in random order. Conversely, the category A exceptions followed the following constraint: item A_2 was presented strictly before item A_5. Figure 2.11 (on the top) shows an example of block following a rule-based order (categories were blocked).

Again, the *similarity-based* order is designed to maximize the similarity between contiguous examples. The first item was randomly selected while subsequent objects were randomly selected among those maximizing the similarity with immediately presented stimuli. Similarly as in Experiment I, similarity between items was computed by counting

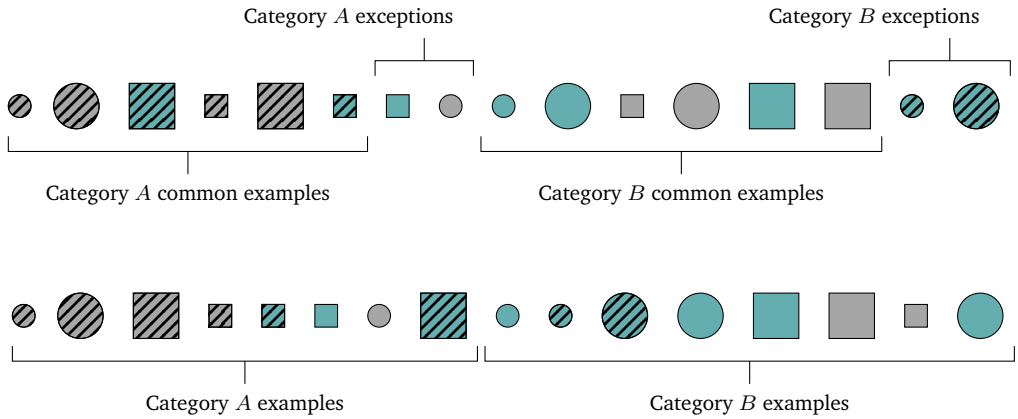


Figure 2.11 – *Examples of blocks following rule-based (on the top) and similarity-based (on the bottom) within-category presentation order by using stimuli from Experiment II. In the rule-based order, the stimuli were ordered following a «principal rule plus exceptions» structure. Conversely, the similarity-based order is designed to maximize the similarity between consecutive stimuli.*

the number of common features between stimuli. Figure 2.11 (on the bottom) shows an example of a block following a similarity-based order (categories were blocked).

Presentation across blocks. In both the Random-Constant and Blocked-Constant contexts a constant across-blocks presentation was considered. However, in the Blocked-Constant context, across-blocks order was manipulated only in odd blocks (even blocks were random blocks).

Conversely, in the Random-Variable context a variable across-blocks presentation was considered. A summary of the different types of orders manipulated in Experiment II as a function of the contexts is given in Figure 2.12.

Stop criterion. The unique learning phase was completed when one of the two following conditions was satisfied (these conditions are equivalent to those described in Experiment I):

i. A sequence of 4×16 consecutive correct responses was given, i.e. 4 consecutive correct blocks of 16 stimuli (in the Blocked-Constant context only odd blocks were considered).

ii. Two sequences of respectively $2 \times 16 + 1$ and 2×16 consecutive correct responses were given, i.e. 2 consecutive correct blocks plus one correct stimulus and 2 consecutive correct blocks. However, between the two sequences, no more than two incorrect responses in a row were approved. Again in the Blocked-Constant context only odd blocks were considered.

Procedure and feedback. There were no warmup session. However, participants were briefly instructed before the task began. Participants were individually tested on a single $12_{4[8]}$ concept. The categorization task was computer-driven and consisted of a single one-hour session (including briefing and debriefing). Participants sat approximately 60 cm from the computer screen on which stimuli were presented.

Stimuli were presented one at a time in the upper part of the screen. Participants sorted the stimuli in one of the two categories by means of two keys. Category A was associated to the up key and was depicted as a school bag located at the top right side of the screen (to match the up key). Category B was associated to the down key and was depicted as a trash located at the bottom right side of the screen (to match the down key).

When the stimulus was presented on the screen, participants were given maximum 8 s to sort it in one of the two categories by means of the response keys. Each time a response key was pressed, the corresponding picture was displayed for 2 s, while the opposite picture disappeared for 2 s. Simultaneously, feedback indicating a correct or incorrect classification of the stimulus was shown at the bottom of the screen for 2 s. If the response was given too late, a "too late" message was displayed on the screen for 2 s. The two category pictures reappeared when a new stimulus was presented.

However, in odd block of the Blocked-Constant context, a slightly different procedure was adopted. When a stimulus was presented, the correct category label (i.e. "school bag" or "trash"), as long as the corresponding category picture, were displayed for 1 s. Simultaneously, the opposite category picture disappeared from the screen for 1 s. This procedure was followed by a confirmation phase in which participants had to press the correct response key. After the key was pressed, feedback indicating a correct or incorrect classification was given at the bottom of the screen for 2 s.

In order to encourage learning, a progress bar representing the score of the participants was displayed at the bottom of the screen. The progress bar was composed of 4×16 empty boxes, which were filled as participants collected points. The participants scored one points each time a correct response was given (for the third context only correct

responses given in odd blocks were taken into account). The progress bar was reset to zero every time an incorrect response was given. An exception to the latter rule was made when participants filled at least half plus one boxes of the progress bar. In this case, every time participants gave an incorrect response (and only one consecutive incorrect answer) the progress bar was reset to the half (i.e. the 2×16-th box). If the response was given too late, participants lost three points on the progress bar.

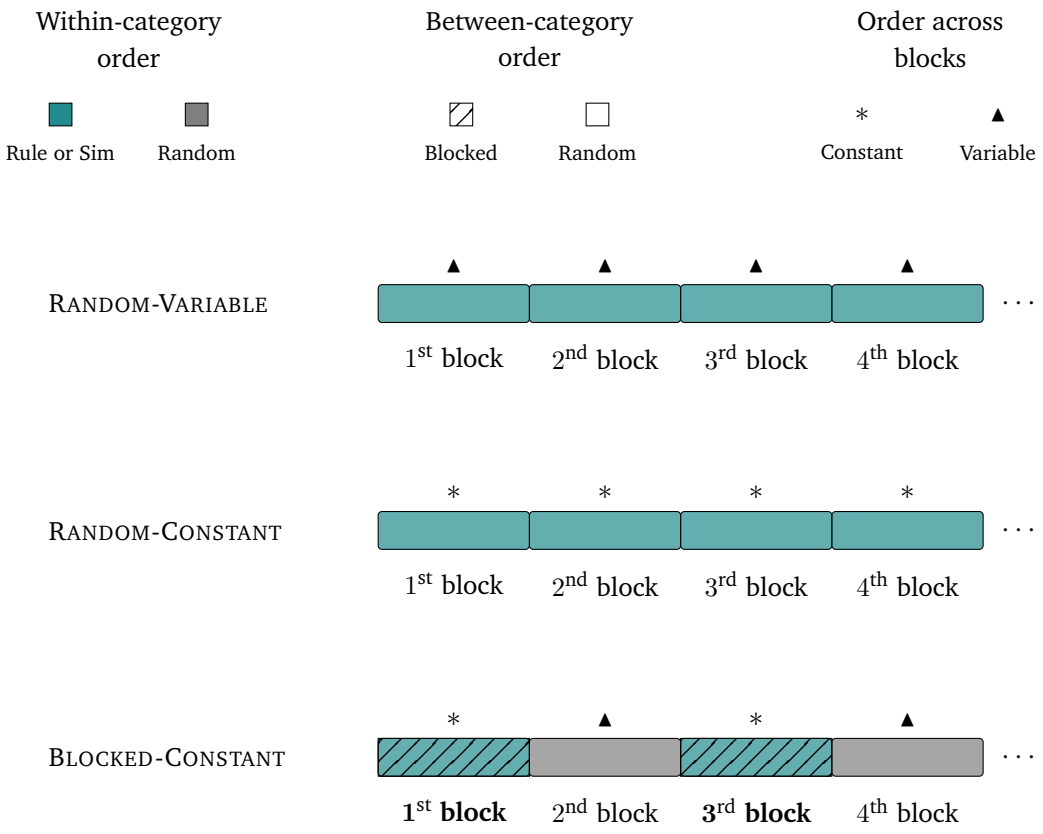


Figure 2.12 – *Summary of the order manipulation in Experiment II as a function of the contexts (Random-Variable, Random-Constant, and Blocked-Constant). Color indicates the within-category presentation order: blue for rule-based or similarity based orders, and gray for random order. Filling pattern indicates the between-category presentation order: striped for blocking, and plain for random. Finally, symbols represent the presentation across blocks: a star for a constant presentation and a triangle for a variable presentation. Bold style is used to indicate that feedback was given both before and after the classification trial.*

TO SUM UP *Experiment II: Data Collection*

Experiment II was designed to investigate the similarity- and rule-based orders in three different contexts: when categories are randomly alternated and a variable across-blocks presentation is considered (Random-Variable context); when categories are randomly alternated and a constant across-blocks presentation is considered (Random-Constant context); and when categories are blocked and a constant across-blocks presentation is considered (Blocked-Constant context).

Participants. There were 68, 22, and 46 participants in respectively the R-V, the R-C, and the B-C contexts.

Phases. The experiment was composed of a unique supervised learning phase.

Categories. Participants were tested on a single $12_{4[8]}$ concept [Fel03]. This concept is composed of 16 items placed on a 4-cube, 8 items belonging to category A and 8 items belonging to category B.

Items. Items varied along four Boolean dimensions: *shape, color, size* and *filling pattern*. Each dimension was not instantiated by the same physical feature.

Between-category presentation order. In the Random-Variable and Random-Constant contexts, categories were randomly alternated. In the Blocked-Constant context, blocks where categories were strictly blocked were alternated with random blocks.

Within-category presentation order. In all contexts, both the rule- and similarity-based orders were used. In the rule-based order, stimuli obeying the principal rule were presented strictly before the exceptions. Conversely, the similarity-based order was designed to maximize the similarity between successive stimuli.

Presentation across blocks. In the Random-Constant and Blocked-Constant contexts, a constant across-blocks presentation was considered. Conversely, in the Random-Variable context, a variable across-blocks presentation was considered.

Stop criterion. The learning phase ended when participants reached the learning criterion, meaning they had to correctly classify two consecutive blocks twice without making two or more mistakes in a row between the two sequences.

Procedure. Each participant was individually tested on a computer-driven task. In all the contexts feedback was given after the classification trial. Only in the odd blocks of the Blocked-Constant context, feedback was additionally given before the classification trial.

2.2.2 Within-Category Order: Rule-based vs. Similarity-based

This subsection is focused on investigating how within-category order (rule-based vs. similarity-based) influences category learning in three contexts (Random-Variable, Random-Constant, and Blocked-Constant). The analysis follows the same path used in Experiment I. We firstly consider unsuccessful participants, then we analyze a set of relevant times, and finally we focus on the proportion of correct responses.

Unsuccessful Participants

As in Experiment I, the aim is to determine whether the number of unsuccessful participants is related with the within-category order.

Fisher's exact test of independence. Table 2.4 shows the number of successful and unsuccessful participants as a function of both the within-category order and the context (Random-Variable, Random-Constant, and Blocked-Constant). A visual representation of Table 2.4 is provided in Figure 2.13. In all contexts, the number of unsuccessful participants was higher in the rule-based order as compared to the similarity-based order. A Fisher's exact test of independence was performed to determine whether this dependency was statistically significant. The test was not significant, with a p-value of 0.46, 0.48, and 0.28 for, respectively, the Random-Variable, Random-Constant, and Blocked-Constant contexts.

We additionally ran a Fisher's exact test of independence in which all three contexts were grouped together. Again, the test was not significant (p-value=0.09). Since the power of a test is influenced by the number of individuals and since the p-value of 0.09 showed an effect, it could be appropriate to assess the relation between number of successful participants and within-category order on a larger sample.

	Successful	Unsuccessful	Total
RANDOM-VARIABLE			
Rule-based	21	13	34
Similarity-based	17	17	34
Total	38	30	68
RANDOM-CONSTANT			
Rule-based	11	0	11
Similarity-based	9	2	11
Total	20	2	22
BLOCKED-CONSTANT			
Rule-based	20	3	23
Similarity-based	16	7	23
Total	36	10	46

Table 2.4 – *Number of successful and unsuccessful participants as a function of both the within-category order (rule-based vs. similarity-based) and the contexts (Random-Variable, Random-Constant, and Blocked-Constant).*

Conclusion: The Fisher's exact test of independence showed the number of unsuccessful participants was not related to the within-category order.

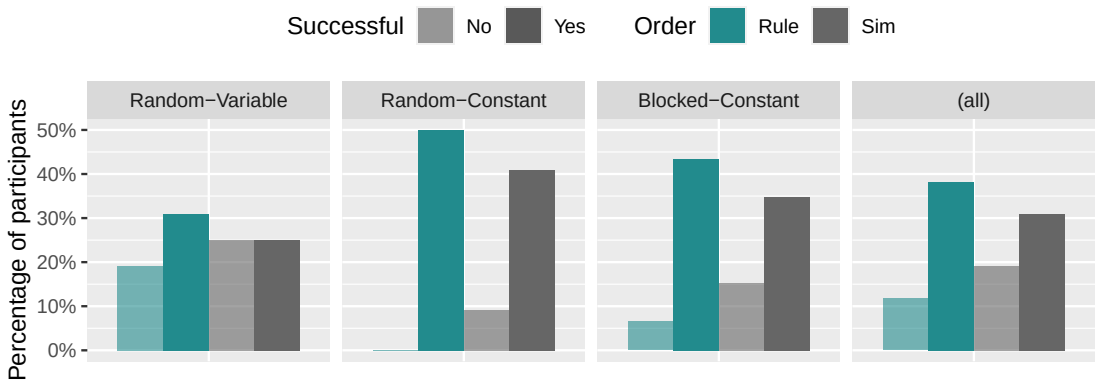


Figure 2.13 – *Percentage of successful and unsuccessful participants as a function of both the within-category order (rule-based vs. similarity-based) and the contexts (Random-Variable, Random-Constant, and Blocked-Constant).*

Relevant Times to Estimate Learning

The analysis on relevant times aims to determine whether learning was faster in the rule-based order as compared to the similarity-based order. As in Experiment I, both classic methods and survival analysis techniques were employed. In view of the high number of unsuccessful participants, the survival techniques provided added value.

Wilcoxon-Mann-Whitney test. The relevant times on which the Wilcoxon-Mann-Whitney test was performed were the same as in Experiment I (i.e., the Ending time, the Learning time, the First time 100%, the First time 75%, and the Never under 60%; see Subsection 2.1.2 for a description). As in Experiment I, the relevant times were computed in terms of stimuli for best accuracy and the one-sided Wilcoxon-Mann-Whitney test was performed (the null hypothesis assumed that rule-based participants had greater relevant times than similarity-based participants). Figure 2.14 shows the average relevant times as a function of both the within-category order and the context. The relevant times of rule-based participants are (on average) smaller than those of similarity-based participants, except for the Learning time and First time 100% of the Random-Variable context. The significance of the one-sided Wilcoxon-Mann-Whitney test are equally showed in Figure 2.14 (see Table 2.2 to map symbols with p-value ranges).

In the Random-Variable context, the test found no difference between the relevant times of the two orders. In the Random-Constant context, all tests were significant, showing that learning was faster in the rule-based order. Finally, in the Blocked-Constant context, the test was significant only in the Ending time and First times 75%. The number of participants that was removed from the tests as well as the specific p-values of the tests are showed in Table 2.5.

Although the test was overall significant, a large number of participants was removed beforehand (i.e., all the participants who did not reach the selected relevant times), producing a loss of information. A further investigation by means of survival techniques is thus valuable for exploring the effects of within-category order without removing participants.

Conclusion: the test was overall significant for both the Random-Constant and Blocked-Constant contexts. However, the removing of a large number of participants represented a considerable downside of the method.

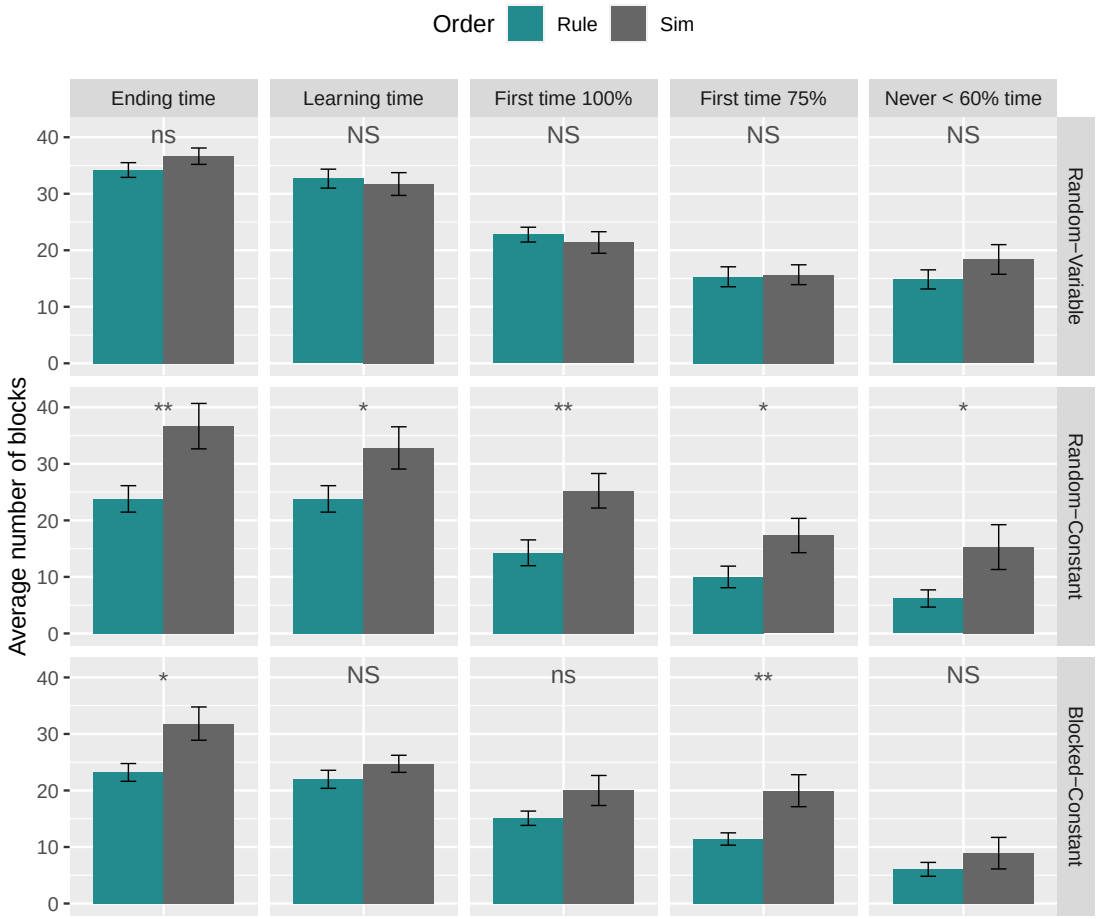


Figure 2.14 – *Average relevant times as a function of both the within-category order (rule-based vs. similarity-based) and the contexts (Random-Variable, Random-Constant, and Blocked-Constant). Both star and "ns" symbols represent the significance of the Wilcoxon-Mann-Whitney test (see Table 2.2 to map symbols with p-value ranges).*

Kaplan–Meier survival curves and Log-Rank test. As we previously saw, the main advantage provided by survival techniques consists in avoiding to remove unsuccessful participants. Since a large number of participants did not reach the learning criterion, this advantage suits our case particularly well.

Figure 2.15 shows the survival curves estimated with the Kaplan-Meier estimator as a function of both the within-category order and the context. We recall that survival curves represents the probability that an individual has not yet met the learning criterion at a certain time. In the Random-Variable context, the rule-based and similarity-based survival curves were close. Conversely, in both the Random-Constant and Blocked-Constant

# Removed subjects	Ending time	Learning time	First Time 100%	First Time 75%	Never <60% time
R-V	-	30	15	6	4
R-C	-	2	-	-	-
B-C	-	10	6	3	2

P-value	Ending time	Learning time	First Time 100%	First Time 75%	Never <60% time
R-V	0.08	0.58	0.72	0.45	0.33
R-C	0.006 **	0.02 *	0.005 **	0.02 *	0.02 *
B-C	0.01 *	0.1	0.06	0.002 **	0.46

Table 2.5 – *Number of removed participants and p-values of the one-sided Wilcoxon-Mann-Whitney test as a function of both the relevant time and the contexts (Random-Variable, Random-Constant and Blocked-Constant).*

contexts, the rule-based survival curve remained under the similarity-based survival curve. A log-rank test was performed to assess the difference between the rule- and similarity-based survival curves. The p-values of the long-rank test were 0.1, 0.005, and 0.04 for, respectively, the Random-Variable, the Random-Constant and the Blocked-Constant contexts. Therefore, in the Random-Constant and Blocked-Constant contexts, rule-based participants had smaller probability to not reach the learning criterion as compared to similarity-based participants.

Conclusion: the log-rank test performed on the Kaplan-Meier survival curves was significant in the Random-Constant and Blocked-Constant contexts, meaning that the rule-based order facilitated the reaching of the learning criterion.

Cox proportional-hazards model. As in Experiment I, an alternative survival analysis technique is represented by the Cox model, whose main advantage is to express the hazard probability (i.e., the probability that a participant reaches the learning criterion at a certain time) as a function of several variable (i.e., the within-category order).

The results of the Cox model are shown in Figure 2.16. The rule-based condition is the reference condition and, consequently, its hazard ratio was 1. The similarity-based

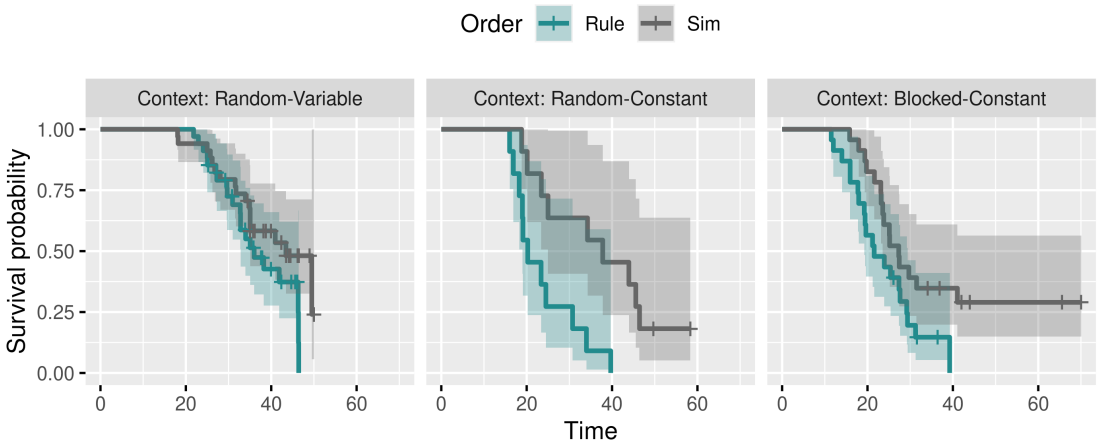


Figure 2.15 – *Illustration of the Kaplan-Meier survival curves of Experiment II as a function of both the within-category order and the context. Survival curves represent the probability that a participant has not yet reached the learning criterion at a certain time. The transparent areas around the survival curves represent the 95% confidence intervals.*

condition is the opposite condition and its hazard ratio varied across the contexts: 0.59 for the Random-Variable context, 0.25 for the Random-Constant context, and 0.5 for the Blocked-Constant context.

However, these results are meaningless if not associated with the test assessing the relevance of the model. The three tests (likelihood-ratio test, Wald test, and log-rank test) assessing the relevance of the model were only significant for the Random-Constant and Blocked-Constant contexts (p-values of 0.007, 0.009, and 0.005 for, respectively, the likelihood-ratio test, the Wald test, and the log-rank test in the Random-Constant context; p-values of 0.04 for all three tests in the Blocked-Constant context). Therefore, in the Random-Constant and Blocked-Constant contexts, similarity-based participants had smaller probability to reach the learning criterion as compared to rule-based participants.

Conclusion: the analysis on the Cox model was significant only in the Random-Constant and Blocked-Constant contexts, meaning that (in these contexts) rule-based order facilitated the reaching of the learning criterion.

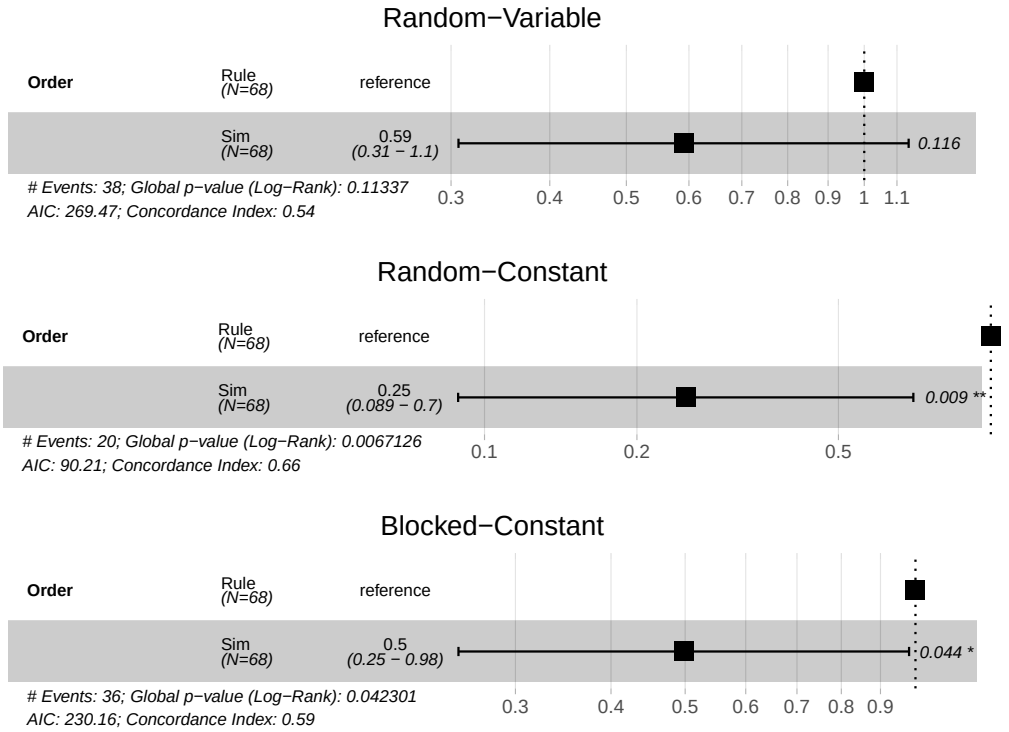


Figure 2.16 – *Results of the Cox proportional-hazards model of Experiment II as a function of both the within-category order and the contexts. The rule-based condition is the reference condition and, consequently, its hazard ratio is always equal to 1. The similarity-based condition is the opposite condition and its hazard ratio is displayed below the word "reference". The numbers within the brackets below the hazard ratios represent the 95% confidence interval. The numbers on the right side of the graph represent the p-values of the Wald test assessing the significance of the model.*

Proportion of Correct Responses

A last analysis investigated the effects of the within-category order on the evolution of the proportion of correct responses. As in Experiment I, we firstly had to ensure that the duration of the learning phase was the same for every participant. Thus, we only considered the data preceding the limit time (the block at which the fastest unsuccessful participant dropped the experiment) and completed the data until this specific time with 100% correct responses.

Table 2.6 shows the number of blocks that the fastest, the average, the median, and the slowest participant took to end the experiment, as a function of the group of individuals they belong to (successful, unsuccessful, or all together) and the context. By ending the experiment we mean reaching the learning criterion, in the case of successful individuals,

	Successful	Unsuccessful	All together
RANDOM-VARIABLE			
Fastest	18	**25**	18
Average	32	39	35
Median	32	39	35
Slowest	49	50	50
RANDOM-CONSTANT			
Fastest	16	**49**	16
Average	27	54	30
Median	23	54	24
Slowest	46	58	58
BLOCKED-CONSTANT			
Fastest	11	**26**	11
Average	23	43	27
Median	23	39	25
Slowest	41	70	70

Table 2.6 – *Number of blocks that the fastest, the average, the median, and the slowest participants took to end the experiment, depending on the class of individuals they belong to (successful, unsuccessful, or all together) and on the context (Random-Variable, Random-Constant, or Blocked-Constant). By ending the experiment we mean reaching the learning criterion, in the case of successful individuals, or dropping the experiment, in the case of unsuccessful participants. In bold letters we indicate the limit time. The numbers were rounded to their nearest smallest integer.*

or dropping the experiment, in the case of unsuccessful participants. The limit time (i.e., the time until which the data were analyzed) is indicated in bold letters.

In the Random-Variable and Blocked-Constant contexts, the number of blocks of the fastest unsuccessful participant (respectively, 25 and 26) was smaller than the number of blocks of the slowest successful participant (respectively, 49 and 41). Thus, analyzing only the data preceding the limit time, led us to ignore a part of the learning process of both successful and unsuccessful participants.

Percentage of successful participants that:	R-V	R-C	B-C
Met the learning criterion before L	18%	100%	69%
Correctly classified 2 blocks in a row before L	34%	100%	86%
Correctly classified 1 block before L	58%	100%	100%

Table 2.7 – *Percentage of successful individuals that satisfied, before the limit time L (the block at which the fastest unsuccessful participant quit the experiment), one of the three listed condition (meet the learning criterion, correctly classify 2 blocks in a row, and correctly classify 1 block), depending on the sub-experiment.*

Conversely, in the Random-Constant context, the number of blocks of the fastest unsuccessful individual (which is 49) is greater than the number of blocks of the slowest successful individual (which is 46), meaning that only a part of the learning process of the unsuccessful participants is ignored.

Table 2.7 shows the percentage of successful participants that reached one of the three following condition before the limit time L: meet the learning criterion, correctly classify 2 blocks in a row, and correctly classify 1 block. This table warns us about the quantity of information (related with successful participants) that was not considered in the analysis.

For instance, in the Random-Variable context, only 18% of the successful participants reached the learning criterion before the limit time. This means that the last chunk of the learning process of 82% of successful participants is ignored.

Wilcoxon-Mann-Whitney tests with Benjamini-Hochberg correction. Once the data was completed, the one-sided Wilcoxon-Mann-Whitney test was applied to the proportion of correct responses or rule- and similarity-based participants at each block until the limit time (the null hypothesis assumed that the similarity-based participants had an higher proportion of correct responses as compared to the rule-based participants). In the Blocked-Constant context, the tests were only applied to even blocks (in odd blocks feedback were also provided before the classification trial). Since multiple tests were performed, they were corrected by means of the Benjamini-Hochberg procedure.

The p-values of the Wilcoxon-Mann-Whitney test (ordered from the smallest to the largest) are shown in Figure 2.17, as a function of the context. The straight line associated with

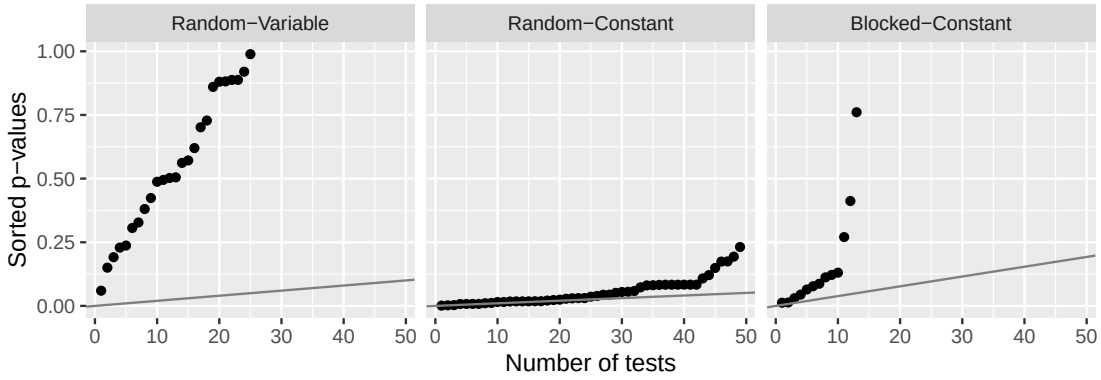


Figure 2.17 – *P-values of the Wilcoxon-Mann-Whitney test performed to compare the proportion of correct responses between rule-based and similarity-based participants of each context of Experiment II. In the Blocked-Constant context, the test was performed only in even blocks. The p-values are ordered from the smallest to the largest. The straight lines are associated with the Benjamini-Hochberg procedure at a significance level of 0.05.*

the Benjamini-Hochberg procedure at a significance level of 0.05 is also plotted. With a significance level of 0.05, all tests of all contexts were accepted.

However, in the Random-Constant context the p-values were close to the straight line. Indeed, by considering a significance level of 0.054, 17 tests over 49 were rejected (35%). The smallest significance level with which at least one test is rejected was 0.95, 0.053, and 0.087 for respectively the Random-Variable, the Random-Constant, and the Blocked-Constant contexts.

The fact that *i)* the power of a test is influenced by the number of individuals, *ii)* the multiple tests correction tends to penalize the rejection, and *iii)* the p-value of the Blocked-Constant context was small, made us think that it could be appropriate to performe the test on a larger sample.

Finally, Figure 2.18 shows the average proportion of correct responses, as a function of both the within-category order and the contexts. The asterisk symbols denote the blocks in which the test was rejected by the Benjamini-Hochberg procedure with a significance level of 0.054.

Conclusion: the Wilcoxon-Mann-Whitney tests with the Benjamini-Hochberg correction were significant only in the Random-Constant context (17 tests over 49 were rejected). Therefore, in this context, the evolution of the proportion of correct responses was higher in the rule-based order as compared to the similarity-based order.

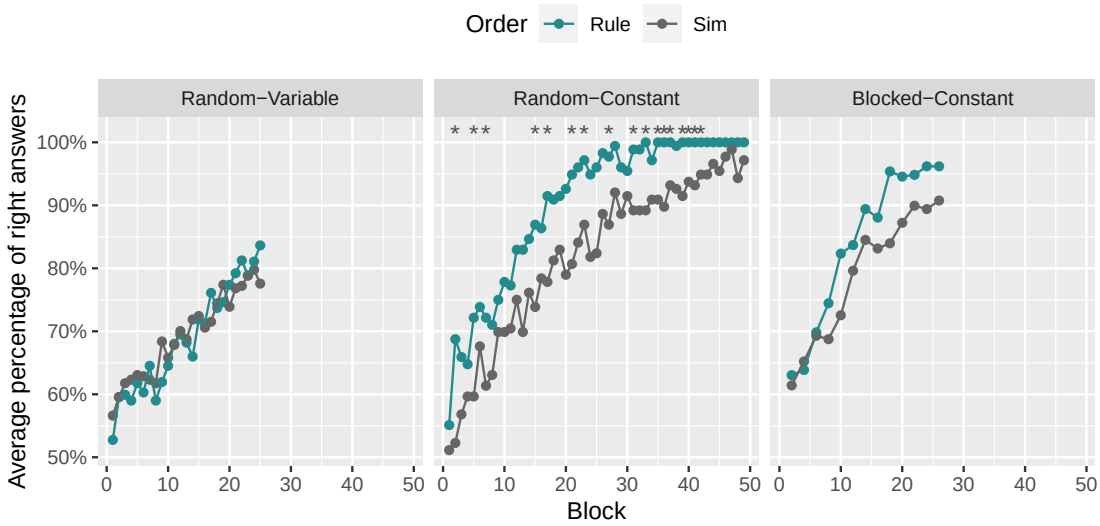


Figure 2.18 – *Average proportion of correct responses in Experiment II, as a function of the within-category order and the contexts. In the Blocked-Constant context, the proportion of correct responses was computed only for even blocks. The asterisk symbol denotes the blocks that have been rejected by the Benjamini-Hochberg procedure with a significance level of 0.054.*

TO SUM UP *Experiment II: Rule-based vs. Similarity-based*

The aim was to investigate the effects of within-category order (rule-based vs. similarity-based) on category learning. The analysis was organized in three parts: *i)* determining whether the number of unsuccessful participants was related to the within-category order, *ii)* analyzing a set of relevant times to compare the learning speed of rule- and similarity-based participants, and *iii)* comparing the evolution of correct responses of rule- and similarity-based participants.

Unsuccessful Participants

Fisher's exact test of independence (not significant). A Fisher's exact test of independence was performed on the number of successful and unsuccessful participants of each context (and of the overall dataset). Although the number of unsuccessful participants was always higher in the similarity-based order as compared to the rule-based order, the tests were not significant.

Relevant Times to Estimate Learning

Wilcoxon-Mann-Whitney test (significant in Random-Constant and Blocked-Constant). Five relevant times were analyzed: the Ending time (the time at which participants ended the learning phase), the Learning time (the time at which successful participants met the learning criterion), the First time 100% (the first time at which participants correctly classify a block), the First time 75% (the first time at which participants correctly classify 75% of items in a block), and the Never time 60% (the time starting from which participants correctly classify at least 60% of items in each block). A one-sided Wilcoxon-Mann-Whitney test was performed to compare the previous relevant times of rule-based and similarity-based participants. The test was overall significant only for the Random-Constant and Blocked-Constant contexts, showing faster learning in the rule-based order.

Kaplan-Meier survival curves and Log-Rank test (significant in Random-Constant and Blocked-Constant). The survival probability of rule- and similarity-based participants was estimated by means of the Kaplan-Meier estimator. The two survival curves were then compared using the log-rank test. The test was significant only in the Random-Constant and Blocked-Constant contexts, showing that the rule-based order facilitated the reaching of the learning criterion.

Cox proportional-hazards model (significant in Random-Constant and Blocked-Constant). The Cox model was used to express the hazard probability as a function of the within-category order. The relevance of the model was then evaluated by means of three tests (likelihood-ratio test, Wald test, and log-rank test). The model was relevant only in the Random-Constant and Blocked-Constant contexts, showing that the similarity-based order reduced the probability to reach the learning criterion.

Proportion of Correct Responses

The dataset was analyzed until the limit time (the block at which the fastest unsuccessful participant dropped the experiment) and completed with 100% of correct responses, allowing a normalization of the length of the learning phase.

Wilcoxon-Mann-Whitney tests with Benjamini-Hochberg correction (significant in Random-Constant)**.** A Wilcoxon-Mann-Whitney test was performed to compare the proportion of correct responses of rule- and similarity-based participants at each block. The result of the tests was then corrected using the Benjamini-Hochberg procedure. The Benjamini-Hochberg procedure at a significance level of 0.054 rejected 17 tests over 49 (35%) in the Random-Constant context, showing that the rule-based order facilitated learning during the entire process. In the other contexts, no effect of the within-category order was detected.

2.2.3 Contexts Comparison: Random-Variable vs. Random-Constant vs. Blocked-Constant

The aim here is to investigate whether different contexts (Random-Variable, Random-Constant, and Blocked-Constant) influenced the learning speed. However, our dataset does not allow us to determine in a conclusive manner why specific types of order facilitated learning. For instance, if participants learned faster in the Random-Constant context rather than the Random-Variable context, we cannot conclude that a constant across-blocks presentation facilitated learning. Indeed, participants could have memorized (consciously or not) the sequences of responses of the constant block. They thus would have scored 100% of correct responses ignoring the stimuli, or at least, relying on fewer information regarding stimulus features.

The argument is similar for attempting to compare Random-Constant and Blocked-Constant, not mentioning that the insertion of random blocks in Blocked-Constant makes this comparison even more unfavorable. A transfer phase would have been ideal to determine the source of the advantage, however a different set of categories should have been selected. Future experiments will be conducted to further explore the effects of different types of order on learning and transfer performance. The analysis follows a

	R-V	R-C	B-C
Successful participants	56%	91%	78%
Unsuccessful participants	44%	9%	22%

Table 2.8 – *Percentage of successful and unsuccessful participants for each one of the sub-experiments. We refer the reader to Table 2.4 for the effective number of successful and unsuccessful participants.*

similar path to that previously used. We firstly consider unsuccessful participants and then analyze a set of relevant times. The comparison of the proportion of correct responses was avoided because of the large disparities in the length of the learning phase.

Unsuccessful Participants

Again, the aim is to determine whether the number of the unsuccessful participants is related to the context (Random-Variable, Random-Constant, and Blocked-Constant).

Fisher's exact test of independence. Table 2.8 shows the percentage of successful and unsuccessful participants for each context. The choice to show the percentage rather than the number of participants is motivated by the disparities regarding the total number of participants (see Table 2.4 to look at the number of participants). A Fisher's exact test of independence was performed to determine whether the participants' outcome (successful or unsuccessful) was influenced by the context. The test showed a strong association between the two variables (p-value=0.002).

Conclusion: the Fisher's exact test of independence showed that the number of unsuccessful participants is strongly dependent from the context of the experiment (Random-Variable, Random-Constant, and Blocked-Constant).

Ending Time

Analyzing the time at which participants ended the experiment aims to compare the learning speed among different context. Again, both classic methods and survival analysis techniques were employed.

Pairs of contexts	R-V vs. R-C	R-C vs. B-C	R-V vs. B-C
P-value	$0.03 *$	0.1	$< 0.001 * * *$

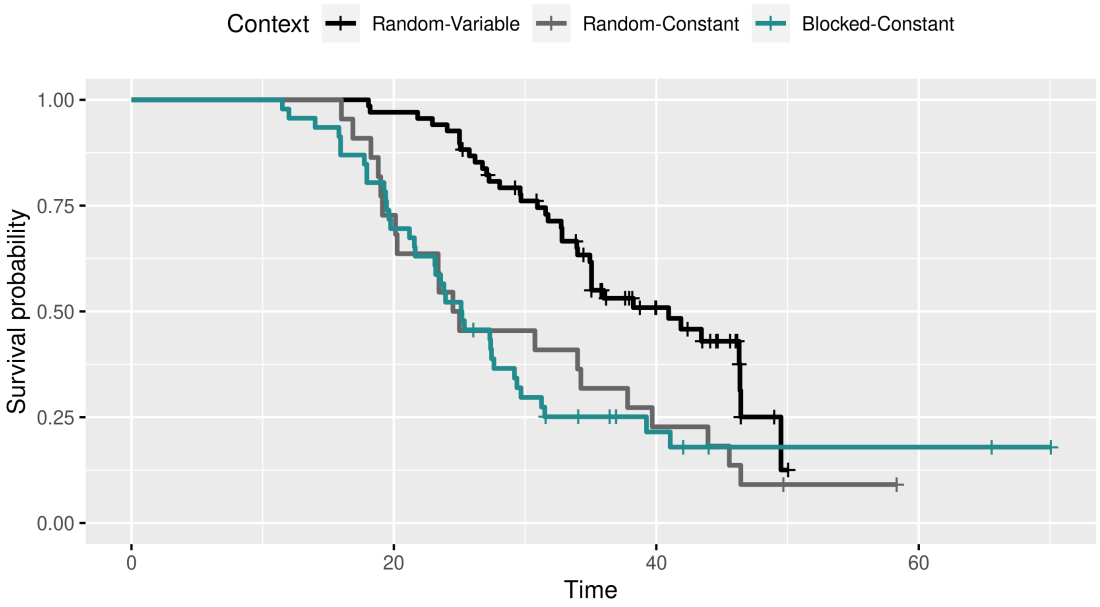
Table 2.9 – *P-values of the Wilcoxon-Mann-Whitney test for each pair of contexts (Random-Variable vs. Random-Constant, Random-Constant vs. Blocked-Constant, Random-Variable vs. Blocked-Constant). The null hypothesis for the pairs Random-Variable vs. Random-Constant, Random-Constant vs. Blocked-Constant and Random-Variable vs. Blocked-Constant assumed that learning was faster in, respectively, the Random-Variable, the Random-Constant, and the Random-Variable contexts.*

Wilcoxon-Mann-Whitney test. Unsuccessful participants were removed from the analysis. This ensured that the ending time corresponded to the time at which participants met the learning criterion. A Wilcoxon-Mann-Whitney test was run to compare the participants' ending times of each pair of contexts (i.e., Random-Variable vs. Random-Constant, Random-Constant vs. Blocked-Constant, Random-Variable vs. Blocked-Constant). Each pair of contexts was characterized by a specific null hypothesis. For the pair Random-Variable vs. Random-Constant, the null hypothesis assumed that learning was slower in the Random-Constant context as compared to the Random-Variable context. For the pair Random-Constant vs. Blocked-Constant, the null hypothesis assumed that learning was faster in the Random-Constant as compared to the Blocked-Constant context. Finally, for the pair Random-Variable vs. Blocked-Constant, the null hypothesis assumed that learning was faster in the Random-Variable as compared to the Blocked-Constant context. Again, the ending time was expressed in terms of stimuli for best accuracy.

The p-values of the Wilcoxon-Mann-Whitney test for each pair of contexts (Random-Variable vs. Random-Constant, Random-Constant vs. Blocked, and Random-Variable vs. Blocked-Constant) are showed in Table 2.9. The test was only significant for the pairs Random-Variable vs. Random-Constant and Random-Variable vs. Blocked-Constant, showing that learning was faster in the Random-Constant and Blocked-Constant contexts as compared to the Random-Variable context.

Conclusion: Participants in the Random-Constant and Blocked-Constant contexts learned faster as compared to the Random-Variable context.

Kaplan–Meier survival curves and Log-Rank test. Again, survival techniques allowed us to avoid the removing of unsuccessful participants and improved the previous analysis. Figure 2.19 (on the top) shows the survival curves of each of the contexts (Random-Variable, Random-Constant, and Blocked-Constant). We recall that a survival curve



Pairs of contexts	R-V vs. R-C	R-C vs. B-C	R-V vs. B-C
P-value	0.005 **	0.9	< 0.001 ***

Figure 2.19 – *Survival curves and p-values of the log-rank test. On the top, Kaplan-Meier survival curves of each context (Random-Variable, Random-Constant, and Blocked-Constant) of Experiment II. On the bottom, the p-values of the log-rank test for each pair of contexts.*

expresses the probability that a participant has not yet reached the learning criterion at a certain time. A log-rank test was performed to assess the distance between each pair of survival curves (the null hypothesis assumed no difference in survival). The results of the log-rank test are shown in Figure 2.19 (on the bottom). Only the difference between the Random-Variable and Random-Constant survival curves, and that between the Random-Variable and Blocked-Constant survival curves were significant, confirming the previous analysis.

Conclusion: the log-rank test performed on the Kaplan-Meier survival curves showed that both the Random-Constant and Blocked-Constant contexts facilitated the reaching of the learning criterion as compared to the Random-Variable context.

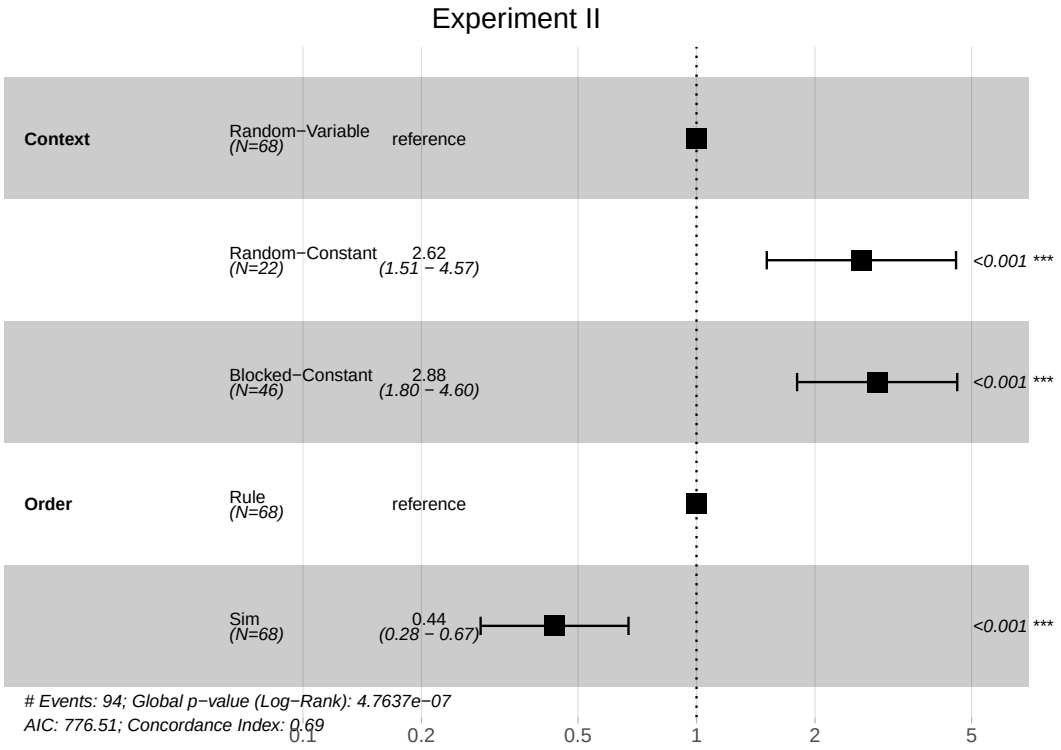


Figure 2.20 – *Results of the Cox model when the hazard probability is expressed as a function of the context (Random-Variable, Random-Constant, and Blocked-Constant) and the within-category presentation order (rule-based and similarity-based). The hazard ratios of the opposite conditions are located below the word "reference". The numbers within the brackets represent the 95% confidence interval. The numbers on the right side of the graph represent the p-values of the Wald test assessing the significance of the model.*

Cox proportional-hazards model. One of the advantage of the Cox model is that it allows us to express the hazard probability as a function of multiple variables. Therefore, we used the Cox model to assess the effects of both the context (Random-Variable, Random-Constant, and Blocked-Constant) and the within-category order (rule-based vs. similarity-based) on learning speed.

The results are shown in Figure 2.20. Regarding the Random-Constant (R-C) context, the combination of a significant p-value (< 0.001) with a hazard ratio equal to 2.62 (> 1) indicates that the Random-Constant context facilitated the reaching of the learning criterion as compared to the Random-Variable context. Similar conclusions are valid for the Blocked-Constant context (p-value < 0.001 and hazard ratio > 1).

Regarding the within-category order, the results showed that the similarity-based order reduced the probability to reach the learning criterion as compared to the rule-based order (p-value < 0.001 and hazard ratio < 1). Moreover, the learning advantage generated by the Random-Constant or Blocked-Constant contexts is almost compensated by the learning disadvantage generated by the similarity-based order. Finally, the tests assessing the overall significance of the model were all significant (p-value < 0.001).

A caveat to this analysis is that the Cox model assumes that the censoring is independent from the studied variables. Yet, the analysis on unsuccessful participants showed that the context influenced the censoring (i.e., the probability to not reach the learning criterion).

Conclusion: the Random-Constant and Blocked-Constant contexts were associated with an increase of probability to reach the learning criterion. Conversely, the similarity-based order was associated with a decrease of probability to reach the learning criterion.

TO SUM UP *Experiment II: Contexts Comparison*

The aim was to investigate the effects of different contexts (i.e., Random-Variable, Random-Constant, and Blocked-Constant) on learning speed. The analysis was organized in two parts: *i)* to determine whether context and number of unsuccessful participants were related, and *ii)* to analyze the duration of the classification task as a function of the context.

Unsuccessful Participants

Fisher's exact test of independence (significant)**.** A Fisher's exact test of independence was performed on the number of successful and unsuccessful participants of each context (i.e., Random-Variable, Random-Constant, and Blocked-Constant). The test showed a strong relation between context and number of unsuccessful participants.

Ending time

Wilcoxon-Mann-Whitney test (significant for Random-Variable vs. Random-Constant and Random-Variable vs. Blocked-Constant)**.** For each pair of contexts, a Wilcoxon-Mann-Whitney test was run to compare the times at which successful participants completed the classification task (the unsuccessful participants were removed from the analysis). The test was significant for the pairs Random-Variable vs. Random-Constant and Random-Variable vs. Blocked-Constant, showing that learning was faster in the Random-Constant and Blocked-Constant contexts as compared to the Random-Variable context.

Kaplan-Meier survival curves and Log-Rank test (significant for Random-Variable vs. Random-Constant and Random-Variable vs. Blocked-Constant)**.** The distance between the Kaplan-Meier survival curves of each pair of contexts was assessed using a log-rank test. The test was significant for the pairs Random-Variable vs. Random-Constant and Random-Variable vs. Blocked, showing that both the Random-Constant and Blocked-Constant contexts facilitated the reaching of the learning criterion as compared to the Random-Variable context.

Cox proportional-hazards model (significant)**.** The Cox model was used to simultaneously assess the effects of both the context and the within-category order. The analysis was significant, showing that *i)* participants in both the Random-Constant and Blocked-Constant contexts had higher probability to successfully complete the task as compared to participants in the Random-Variable context, and *ii)* participants in the similarity-based order had lower probability to meet the learning criterion as compared to participants in the similarity-based order.

2.3 Discussion

The main aim of the present chapter was to investigate the effects of within-category order on learning speed. The statistical analyses that were performed on both Experiment I and II showed three phenomena. Firstly, the rule-based order was not equally beneficial in every studied context. Indeed, in both the Random-Constant and Blocked-Constant contexts, learning was faster in the rule-based order as compared to the similarity-based order. Conversely, in the Random-Variable context, no significant difference was

found between the learning speed of rule- and similarity-based participants. Secondly, the analysis on unsuccessful participants showed that the number of participants who dropped the experiment was not related to the within-category order. Therefore, the disadvantage generated by the similarity-based order did not influence participants' ability to complete the task. Finally, rule-based participants provided better proportion of correct responses across the entire learning phase (as compared to similarity-based participants) in the Random-Constant context exclusively.

In view of these results, the rule-based order has benefited from the Random-Constant and Blocked-Constant contexts. On one hand, the constant blocks in the Random-Constant context might have helped participants to focus their attention toward a limited set of information, or might have induced participants to abstract erroneous rules. In both cases, a rule-based strategy was reinforced, benefiting a rule-based order. On the other hand, the blocked categories in the Blocked-Constant context might have facilitated the detection of a "principal rule plus exceptions" structure, encouraging participants to adopt a rule-based strategy. Indeed, both blocking and rule-based order direct participants' attention toward the similarities within a category, enhancing the probability to abstract the simplest rule.

An additional aim was to investigate whether the learning speed was influenced by different contexts (Random-Variable, Random-Constant, and Blocked-Constant). We found that participants in both the Random-Constant and Blocked-Constant contexts completed the classification task faster than participants in the Random-Variable context. However, our dataset did not allow us to conclusively compare different factors (e.g., constant across-blocks manipulation, blocking, etc.). For instance, faster learning in the Random-Constant context as compared to the Random-Variable context might have been produced by the constant across-blocks manipulation, or by the fact that participants took profit of the repetition of the constant block to memorize the sequence of responses. Future perspectives include a full factorial experiment involving the eight following experimental conditions: rule-based vs. similarity-based types $\times$ interleaved vs. blocked categories $\times$ variable vs. constant across-blocks manipulations. This future experiment will aim to compare the effects of different types of order (blocking, interleaving, across-blocks manipulation, etc.) on learning and transfer performance.

3

Categorization Models

Contents

The aim of the present chapter is to present the categorization models that were used to investigate the cognitive process underlying category learning. This chapter also includes the development of a new exemplar model that extends a reference model in categorization (the GCM) accounting for the order in which stimuli are presented. The structure of this chapter is based on the distinction between transfer and learning models. As we described in the introduction (see "Ability to Learn" in Section 1.4), categorization models can be divided into models that are only suitable for reproducing transfer performance (transfer models) and models that can account for both the learning dynamics and transfer performance (learning models).

Outline of this chapter

Firstly, we describe the mathematical framework shared by the two types of categorization models. Then, we describe the selected transfer models: the Generalized Context Model (GCM), the new Ordinal General Context Model (OGCM), and the Generalized Context Model equipped with a lag mechanism (GCM-Lag). Finally, we present the selected learning models: the Component-Cue model and the Attention Learning COVEring map model (ALCOVE).

3.1 Mathematical Formalization

The mathematical framework shared by transfer and learning models is preceded by the formalization of the canonical categorization experiment. For purposes of clarity, a single participant and a single phase (learning or transfer) are considered. The set of items that are presented to the participant is denoted by E and we assume that there are two categories (A and B) in which stimuli are classified. A categorization task can be described using three layers:

Sequence of stimuli. This is the sequence of stimuli that has been presented to the participant. The sequence of stimuli is denoted by:

$$x^{(1)}, \ldots, x^{(n)},$$

where $x^{(t)}$ represents the t-th stimulus presented to the participant and $x^{(t)} \in E$ for all $t \in \{1, \ldots, n\}$. The sequence of stimuli can be composed of learning items, if the learning phase is considered, or both learning and transfer items, if the transfer phase is considered.

Sequence of responses. This is the sequence of responses given by the participant. Each response corresponds to the category (A or B) in which participant classified the presented stimulus. The sequence of responses is denoted by:

$$y^{(1)}, \ldots, y^{(n)},$$

where $y^{(t)}$ represents the t-th response given by the participant and $y^{(t)} \in \{A, B\}$ for all $t \in \{1, \ldots, n\}$.

Sequence of feedback. This is the sequence of feedback given to the participant after the classification trial. Each feedback corresponds to the category (A or B) of the presented stimulus. The sequence of feedback is denoted by:

$$v^{(1)}, \ldots, v^{(n)},$$

where $v^{(t)}$ represents the t-th feedback given to the participant and $v^{(t)} \in \{A, B\}$ for all $t \in \{1, \ldots, n\}$. If a transfer phase is considered, there is no sequence of feedback (no feedback is provided during the transfer phase).

For practical reasons, a numeric equivalent of the sequence of responses is considered. More precisely, a new sequence is created by replacing the responses A and B in the sequence of responses with, respectively, 1 and 0. The new sequence is denoted by:

$$z^{(1)}, \ldots, z^{(n)}, \qquad \text{such that} \qquad z^{(t)} = \begin{cases} 1 & \text{if } y^{(t)} = A \\ 0 & \text{if } y^{(t)} = B \end{cases},$$

for all $t \in \{2, \ldots, n\}$. The history of the process at time $t \in \{1, \ldots, n\}$ is represented by the sequences of stimuli and feedback until time t and is denoted by:

$$\mathcal{H}_t = \left(x^{(j)}, v^{(j)}\right)_{j=1}^{t}.$$

By convention, we set $\mathcal{H}_0 = \emptyset$. Before describing the mathematical background, a few more notations are needed. Let M be a model, θ its parameters, and $\mathbb{P}_M^\theta\left(A \,|\, x^{(t)}, \mathcal{H}_{t-1}\right)$ the probability of classifying $x^{(t)}$ into category A, knowing $\mathcal{H}_{t-1}$ and given M and θ.

Assumption 3.1. *We assume that there is a sequence of random variables $Z^{(1)}, \ldots, Z^{(n)}$ such that $z^{(t)}$ is a realization of $Z^{(t)}$ and*

$$Z^{(t)} \sim \mathcal{B}\left(\mathbb{P}_M^\theta\left(A \,|\, x^{(t)}, \mathcal{H}_{t-1}\right)\right),$$

where $\mathcal{B}$ is the Bernoulli distribution and $t \in \{1, \ldots, n\}$.

Recall 3.1. If a random variable X follows a Bernoulli distribution of parameter p ($X \sim \mathcal{B}(p)$), then X takes the value 1 with probability p and the value 0 with probability $1 - p$. ⊠

Assumption 3.1 represents our mathematical framework. Regarding transfer models, we additionally assume that the random variables $Z^{(1)}, \ldots, Z^{(n)}$ are independent.

Assumption 3.2. *If M is a transfer model, we additionally assume that:*

$$\mathbb{P}_M^\theta\left(A \,|\, x^{(t)}, \mathcal{H}_{t-1}\right) = \mathbb{P}_M^\theta\left(A \,|\, x^{(t)}\right).$$

In the following sections, the explicit form of the probability of classifying an item into category A (i.e., $\mathbb{P}_M^\theta\left(A \,|\, x^{(t)}, \mathcal{H}_{t-1}\right)$) will be given for each selected models.

3.1.1 Likelihood

In this subsection we present the likelihood of both transfer and learning models. As seen in the introduction, the likelihood measures the goodness-of-fit of the predictions of a model to a dataset, as a function of the parameters of the model. Mathematically speaking, the likelihood is given by the following expression:

$$\mathcal{L}_M\left(z^{(1)}, \ldots, z^{(n)}; \theta\right) = \mathbb{P}_M^\theta\left(Z^{(1)} = z^{(1)}, \ldots, Z^{(n)} = z^{(n)} \,\middle|\, \mathcal{H}_{n-1}\right), \tag{3.1}$$

where M is a model, θ its set of parameters, $z^{(1)}, \ldots, z^{(n)}$ are the participant's responses (the observed values), and $\mathcal{H}_{n-1}$ is the history of the process. Since the likelihood will be extensively used in Chapter 4 and 5, here we show its explicit form.

Transfer Models

If a transfer model is considered, then the independence assumption (Assumption 3.2) can be used to develop Equation 3.1 as follows:

$$\begin{aligned}
&\mathcal{L}_M\left(z^{(1)}, \ldots, z^{(n)}; \theta\right) \\
&\quad = \mathbb{P}_M^\theta\left(Z^{(1)} = z^{(1)}, \ldots, Z^{(n)} = z^{(n)} \,\middle|\, \mathcal{H}_{n-1}\right) \\
&\quad = \prod_{i=1}^{n} \mathbb{P}_M^\theta\left(Z^{(i)} = z^{(i)} \,\middle|\, x^{(i)}\right) \\
&\quad = \prod_{i=1}^{n} \left(\mathbb{P}_M^\theta\left(A \,\middle|\, x^{(i)}\right)\right)^{z^{(i)}} \cdot \left(1 - \mathbb{P}_M^\theta\left(A \,\middle|\, x^{(i)}\right)\right)^{1-z^{(i)}}.
\end{aligned} \tag{3.2}$$

Remark 3.1. The previous equation can be further simplified when the predictions of transfer models only depend on the physical features of the items. The simplification is the following:

$$\mathcal{L}_M\left(z^{(1)},\ldots,z^{(n)};\theta\right)=\prod_{\xi\in E}\left(\mathbb{P}_M^{\theta}\left(A\,|\,\xi\right)\right)^{N_\xi}\cdot\left(1-\mathbb{P}_M\left(A\,|\,\xi\right)\right)^{L-N_\xi},$$

where E is the set of items, N_ξ is the number of times the participant classified item ξ into category A, and L is the number of blocks. ⊠

Learning Models

Conversely, if a learning model is considered, the Bayes' formula can be used to develop Equation 3.1 as follows:

$$\begin{aligned}
&\mathcal{L}_M\left(z^{(1)},\ldots,z^{(n)};\theta\right)\\
&\quad=\mathbb{P}_M^{\theta}\left(Z^{(1)}=z^{(1)},\ldots,Z^{(n)}=z^{(n)}\,\middle|\,\mathcal{H}_{n-1}\right)\\
&\quad=\mathbb{P}_M^{\theta}\left(Z^{(2)}=z^{(2)},\ldots,Z^{(n)}=z^{(n)}\,\middle|\,Z^{(1)}=z^{(1)},\mathcal{H}_{n-1}\right)\cdot\mathbb{P}_M^{\theta}\left(Z^{(1)}=z^{(1)}\,\middle|\,\mathcal{H}_0\right)\\
&\quad=\mathbb{P}_M^{\theta}\left(Z^{(3)}=z^{(3)},\ldots,Z^{(n)}=z^{(n)}\,\middle|\,Z^{(1)}=z^{(1)},Z^{(2)}=z^{(2)},\mathcal{H}_{n-1}\right)\\
&\qquad\qquad\cdot\mathbb{P}_M^{\theta}\left(Z^{(2)}=z^{(2)}\,\middle|\,Z^{(1)}=z^{(1)},\mathcal{H}_1\right)\cdot\mathbb{P}_M^{\theta}\left(Z^{(1)}=z^{(1)}\,\middle|\,\mathcal{H}_0\right)\\
&\quad=\prod_{i=1}^{n}\mathbb{P}_M^{\theta}\left(Z^{(i)}=z^{(i)}\,\middle|\,\bigcap_{1\le j<i}Z^{(j)}=z^{(j)},\mathcal{H}_{i-1}\right).
\end{aligned}\tag{3.3}$$

3.2 Transfer Models

The aim of this section is twofold: *i)* to develop a new extension of the famous Generalized Context Model (GCM), that we called Ordinal General Context Model (OGCM), and *ii)* to describe the transfer models that were used to investigate category transfer.

The selected transfer models are the following: the Generalized Context Model (GCM) [Nos86], which is one of the most popular categorization models; the new Ordinal General Context Model (OGCM), a modification of the GCM that integrates the presentation order; and the Generalized Context Model equipped with a lag mechanism (GCM-Lag), which accounts for sequential effects. Exemplar models were preferred to prototype models

because they generally provide a better account of the data than prototype models [AM93; DG97; NSM17; NS05; NZ02; SM98; SDR00].

The rationale here is the following: although the GCM has extensively proved to provide accurate predictions [NKM92; NSM17; NSM18b; NSM18a; Nos+18; RH05; SN20], it is not sensitive to different presentation orders. Therefore, we developed a modification of the GCM that accounts for the order in which stimuli are presented. This new model, called Ordinal General Context Model (OGCM), was declined in three versions aiming at investigating whether transfer performance was influenced by either *i)* the average order received during learning, or *ii)* the most frequent order received during learning, or *iii)* the order received during transfer. To complete the picture, the Generalized Context Model equipped with a lag mechanism was also considered. The GCM-Lag allowed us to determine whether sequential effects affected participants' performance.

Common Assumptions

All selected transfer models are exemplar models, meaning that they store every distinct occurrences of an item and classify new items as a function of their similarity to the previously stored items. Exemplar models are based on the following assumptions: *i)* items are considered as points of a multidimensional psychological space, *ii)* the previous multidimensional psychological space is equipped with a distance (in this way the distance between two items can be quantified), and *iii)* the similarity between items is defined as a decreasing function of their distance. The first assumption raises the question of how the multidimensional psychological space is estimated. There are three main approaches:

Similarity-scaling plus MultiDimensional Scaling (MDS). The first approach relies on similarity-scaling methods. In this approach, participants are asked to provide similarity judgments among pairs of items. This collection of similarity judgments is then used to generate the psychological space in which items lie through a method called MultiDimensional-Scaling (MDS) [Tor52]. Examples of this approach can be found in [KW78; Nos86; NSM17; Nos+17; Nos+18; SN20; She80] (see also Example 3.1). More recently, MDS representations have been also used to train deep Convolutional Neural Networks (CNNs) to automatically derive psychological representations of new items [SN20].

Considering major features. The second approach consists in using the major features of the items to generate the psychological space. Each major feature corresponds

to a dimension of the space and the different options of each feature are coded by different values. This approach can be applied only with separable-dimension stimuli (see [SHJ61; NP96]; see also Example 3.2).

Dimension ratings. A similar approach consists in collecting direct dimension ratings for the stimuli. The psychological space is then obtained averaging ratings along each of the rated dimensions (see [NSM18a; Swe+91]).

Example 3.1. This example aims to illustrate the MultiDimensional Scaling (MDS) method. Let us consider some of the most famous Italian cities: Florence, Genoa, Milan, Naples, Palermo, Rome, Turin and Venice. We asked members of our laboratory to rate the distance between each pairs of cities. After averaging the ratings, the MDS approach was applied to recover (up to a linear transformation and a re-scaling) the longitude and the latitude of the cities (see Figure 3.2). For sake of simplicity, we fixed the dimension of the space in which points are embedded. However, the multidimensional scaling technique is capable of recovering both the dimension of the space and the coordinates of the points. ⊠

Example 3.2. This example aims to illustrate the second approach to derive the psychological space. Let us consider four objects: a black square, a gray square, a black circle and a gray circle. Since these items are defined by two features (shape and color), they can be embedded in a two-dimensional space where the first dimension corresponds to the shape of the items and the second dimension corresponds to their color. Moreover, specific values can be attributed to the two options of shape (for instance, 0 to the square and 1 to the circle) and to the two options of color (for instance, 0 to black and 1 to gray). Each item has then been associated with a point of a two-dimensional space (see Figure 3.1). ⊠

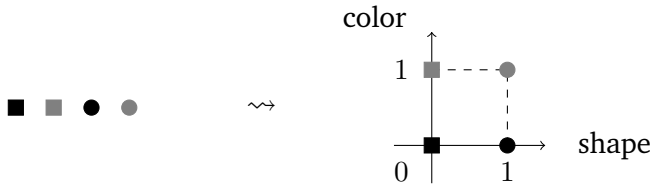


Figure 3.1 – *Illustration of the second approach in which items are associated to a psychological space through their features.*

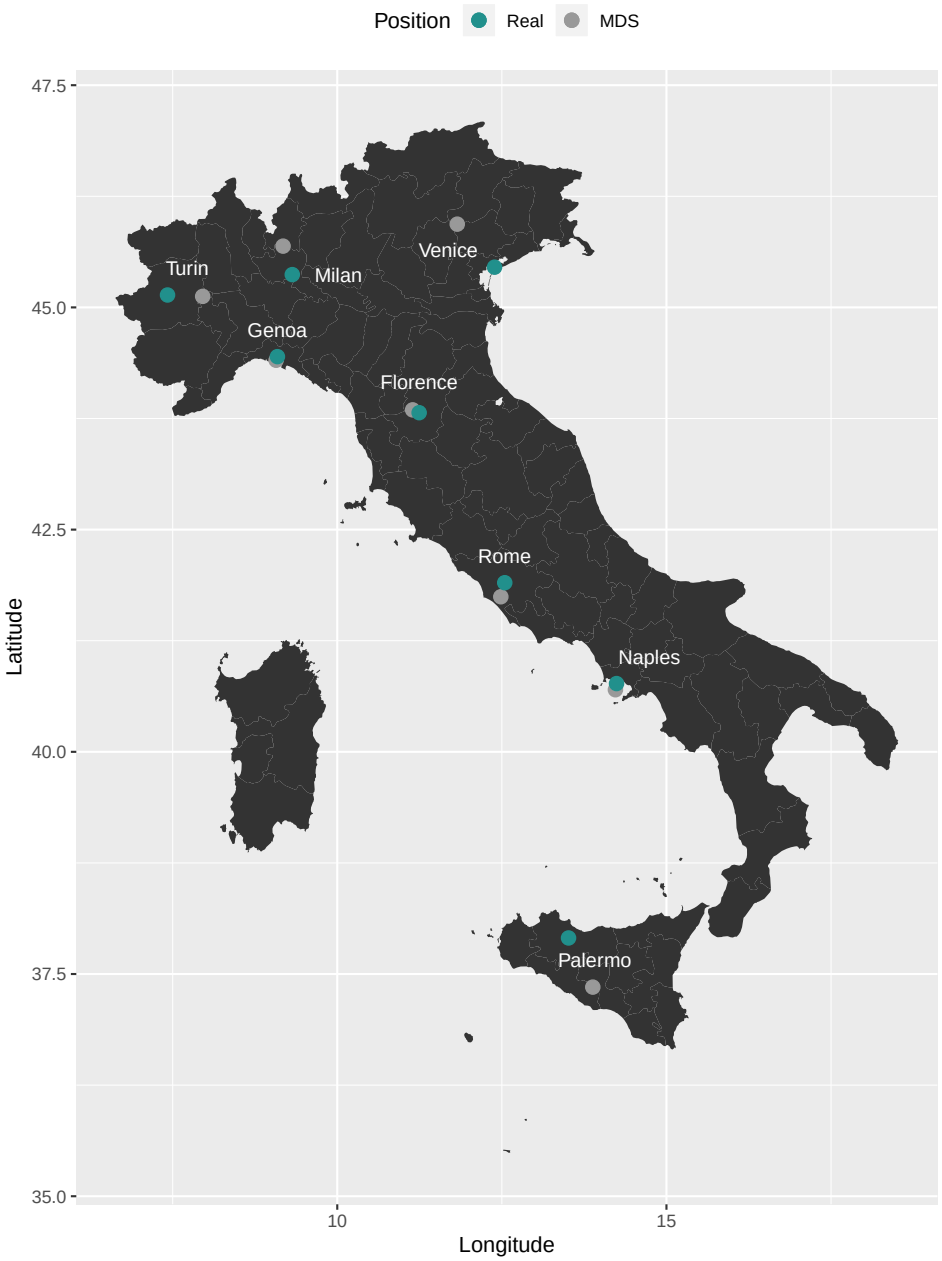


Figure 3.2 – *Application of the MultiDimensional Scaling (MDS) technique to perceived distances of some famous Italian cities. In turquoise blue we can see the real locations of the cities and in gray those recovered with the MDS algorithm. Perceived distances are averaged among 10 individuals.*

3.2.1 Generalized Context Model (GCM)

The Generalized Context Model (GCM) [Nos84; Nos86] was introduced by Nosofsky as a generalization of the context theory of classification developed by Medin and Shaffer [MS78] (see Box 3.1 for further details). The GCM represents a reference point in categorization and has proved to accurately account for transfer performance in a large variety of studies [Nos86; NKM92; NSM17; NSM18b; NSM18a; Nos+18; RH05; RR04; SN20; SM00].

Box 3.1 *Context Theory*

The context theory (also called context model) was developed by Medin and Schaffer in 1978 to account for participants' transfer performance in classification tasks. In what follows, we give a mathematical description of the context model. Let us assume that the psychological space in which stimuli are embedded has dimension $\mathfrak{N}$ and let $\xi = (\xi_i)_{i=1}^{\mathfrak{N}}$ be an item. Moreover, let E_L be the set of learning items, which are classified into two categories (A and B). According to the context theory, the probability of classifying stimulus x as a member of category A (during the transfer phase) is given by:

$$\mathbb{P}(A \,|\, x) = \frac{\sum_{\xi \in A \cap E_L} S(\xi, x)}{\sum_{\xi \in A \cap E_L} S(\xi, x) + \sum_{\xi \in B \cap E_L} S(\xi, x)}, \tag{3.4}$$

where $S(\xi, x)$ represents the similarity between the items ξ and x. The similarity between two items is computed as follows:

$$S(\xi, x) = \prod_{i=1}^{\mathfrak{N}} s_i^{\delta_i(\xi, x)}, \tag{3.5}$$

where s_i (for $i = 1, \ldots, \mathfrak{N}$) are free estimated parameters between 0 and 1, and

$$\delta_i(\xi, x) = \begin{cases} 1 & \text{if } \xi_i \neq x_i \\ 0 & \text{if } \xi_i = x_i. \end{cases}$$

The context model can only be applied to stimuli varying along separable and binary-valued dimensions (see Equation 3.5). Conversely, the GCM generalizes

its application to integral-dimension stimuli as well as non-binary-valued stimuli. The context model has provided good results in numerous classification paradigms [MS78; Med+82; MDM83].

Mathematical Description

In this section, a mathematical description of the GCM is provided. For purposes of clarity, the simplest version is given (its most general version is given in Remark 3.5). The mathematical description is structured in three steps to ensure grater clarity.

Step 1. In this step, the multidimensional psychological space is equipped with a distance (see Section 3.2 to recall how the psychological space is generated).

Step 2. In this step, the distance introduced in Step 1 is used to define the similarity between two items.

Step 3. Finally, the probability of classifying stimuli into category A is defined as a function of their similarity to the stored stimuli.

Step 1. As seen in Section 3.2, the GCM is built on the assumption that items are represented as points in a multidimensional psychological space. Let us assume that this space is of dimension $\mathfrak{N}$ and let $\xi = (\xi_i)_{i=1}^{\mathfrak{N}}$ be an item. The $\mathfrak{N}$-dimensional psychological space is equipped with the following distance:

$$d(\xi, x) = \left[\sum_{i=1}^{\mathfrak{N}} \omega_i \cdot |\xi_i - x_i|^r \right]^{\frac{1}{r}}, \tag{3.6}$$

where ξ and x are two items, r is a positive constant, and ω_i (for all $i = 1, \ldots, \mathfrak{N}$) are freely estimated attention-weight parameters satisfying the following conditions: $\sum_{i=1}^{\mathfrak{N}} \omega_i = 1$ and $\omega_i \geq 0$. The present distance is called the weighted Minkowski distance. The set of attention-weight parameters is denoted by $\omega = (\omega_i)_{i=1}^{\mathfrak{N}}$.

The value r determines the form of the distant metric and is chosen depending on the nature of the items. In experiments involving highly separable-dimension stimuli [Gar74; She64], the value r is usually set equal to 1. Conversely, in situations involving integral-dimension stimuli, the value r is typically set equal to 2. Figure 3.3 (on the top) shows

unit circles with various values of r. Since both Experiment I and II involved highly separable-dimension stimuli, the value of r is set equal to 1.

Remark 3.2. When r is equal to 1, the distance between two points is computed considering the shortest path between the two points generated with vertical and horizontal lines (see Figure 3.3, on the bottom). This distance is called city-block or Manhattan distance. When r is equal to 2, the distance between two points is computed measuring the straight line connecting the two points (see Figure 3.3, on the bottom). This distance is called Euclidean distance. ⊠

The attention-weight parameter ω_i ($i \in \{1, \ldots, \mathfrak{N}\}$) in Equation 3.6 represents the degree of attention that participants gave to dimension i during the classification trials. In other words, the closer the parameter ω_i to 1, the higher the degree of attention given to dimension i. The attention-weight parameters allow the model to stretch the psychological space along the more relevant dimensions and to shrink it along the more irrelevant ones. Let us give an example to better understand the role of the attention-weight parameters.

Example 3.3. Let us take four items placed on $\mathbb{R}^2$ (Figure 3.4, on the left) and let us set the value r equal to 1. Since the psychological space has dimension two, the first attention-weight parameter regulates the attention on the first dimension, while the second attention-weight parameter regulates the attention on the second dimension. The distance between items can be modified selecting different values for the attention-weight parameters (see Figure 3.4, on the right). ⊠

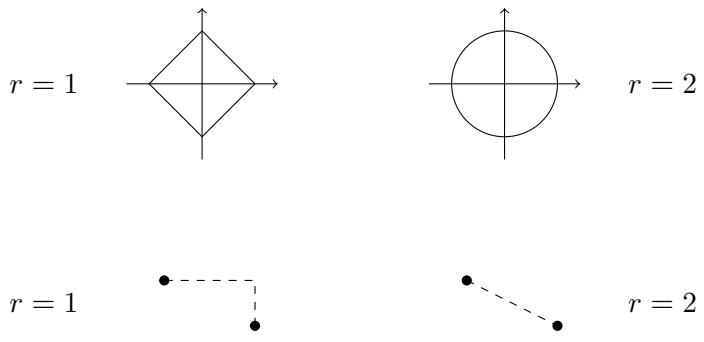


Figure 3.3 – *Variation of the distance, depending on the value of r. On the top, unit circles (i.e. the set of all points that are at the unit distance from the center) with different values of r. The value $r = 1$ yields the city-block distance, while the value $r = 2$ yields the Euclidean distance. On the bottom, the way in which the distance between points is visualized, depending on the value of r.*

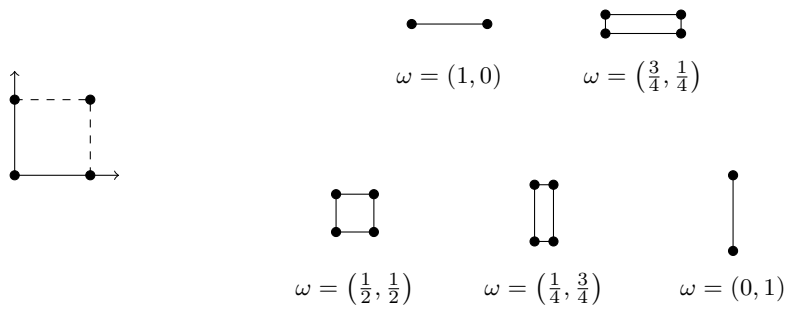


Figure 3.4 – *Effect of the attention-weight parameters on the multidimensional psychological space. On the left, the position of four items on a two-dimension psychological space. On the right, the distance variations due to different attention-weight parameters.*

Step 2. The second step consists in describing the similarity between two items as a function of their distance. According to the GCM, the similarity between items ξ and x is defined as follows:

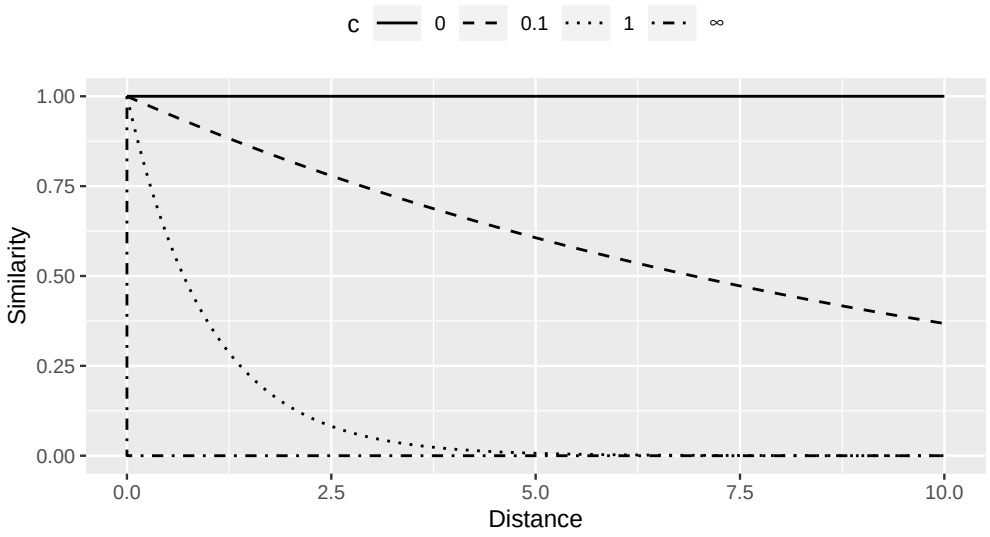
$$S(\xi, x) = e^{-cd(\xi,x)^p}, \tag{3.7}$$

where p is a positive constant and c is a freely estimated sensitive parameter.

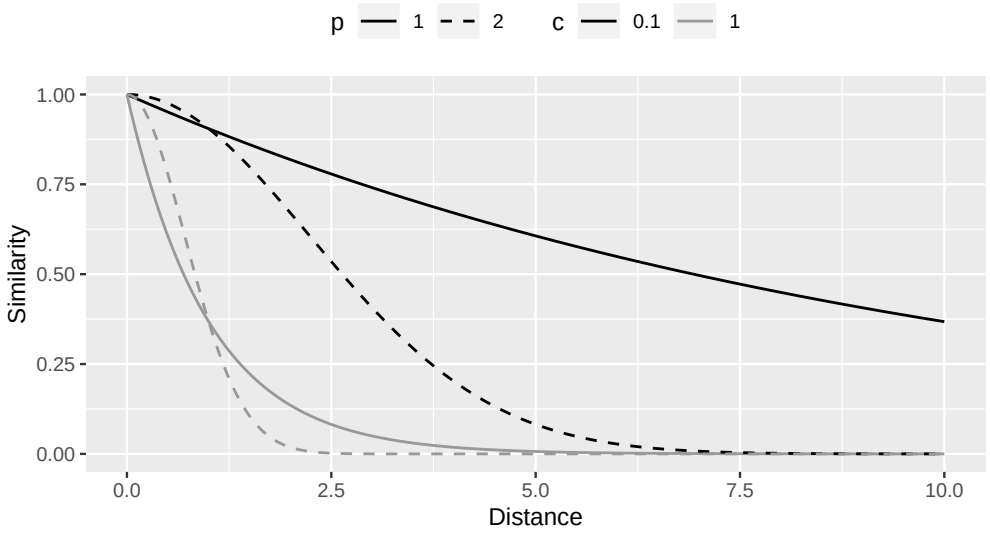
The value c reflects the speed at which similarity decreases with distance (see Figure 3.5, on the top). The greater the value c, the smaller the similarity between close items. In other words, when c is small, close items are perceived as very similar. Conversely, when c is high, close items are perceived as remarkably dissimilar. The extreme cases are when c is equal to 0 or ∞. When c is equal to 0, the similarity between two arbitrary items is equal to 1 (i.e., they are perceived as the same item). On the other hand, when c is equal to ∞, the similarity between two distinct objects is equal to 0 (i.e., all items are perceived as highly dissimilar except for itself). This implies that items are considered to have no similarity between each other, except for themselves.

The value p in Equation 3.7 regulates the shape of the function that relates distance to similarity (see Figure 3.5, on the bottom). When stimuli are highly distinguishable, p is usually set equal to 1 [She87] (exponential relation between distance and similarity). Conversely, when stimuli are not easily distinguishable, p is usually set equal to 2 [Nos85] (Gaussian relation between distance and similarity). Since both Experiment I and II involved distinguishable stimuli, the value of p is set equal to 1.

Remark 3.3. Similarity is a highly context-dependent relation (see [MS78]). For instance, cats and jaguars may be judged as similar in the context of species, but they would



(a) *With p fixed.*



(b) *Both c and p vary.*

Figure 3.5 – *Similarity as a function of the distance with different values of both the sensitive parameter c and the constant p. On the top, the value p is fixed equal to 1. On the bottom, both c and p vary.*

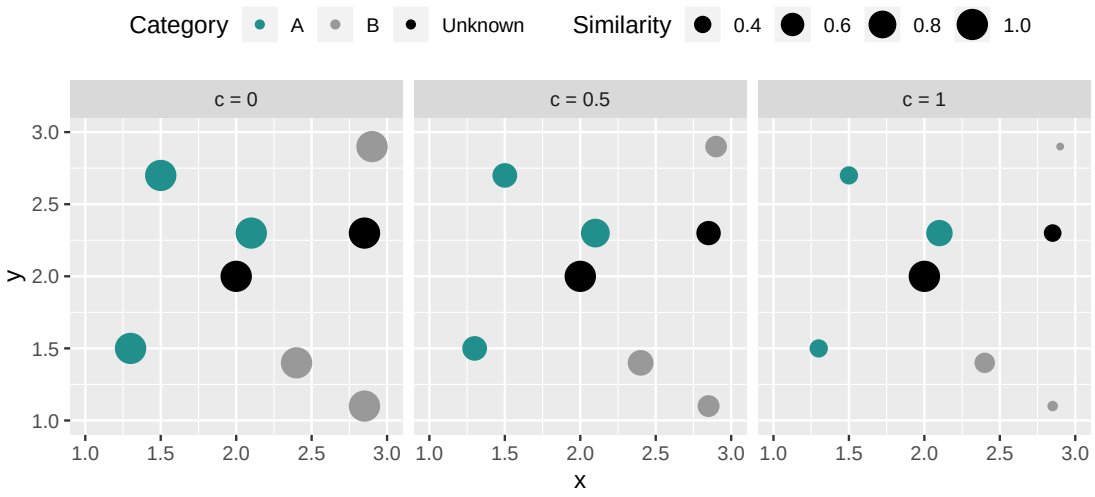


Figure 3.6 – *Impact of the value of the sensitive parameter c on similarity. The colored dots represent the learning items (assigned to one of the two categories A and B), while the black spots represent the transfer items. The size of dots reflects the similarity between the items and the reference item in the middle. The parameters were set as follows: $r = 2$ and $p = 1$.*

be judged as dissimilar in a context that emphasizes danger. In the GCM, this context-dependent nature is shaped by the presence of the attention-weight parameters. ⊠

Example 3.4. This example aims to help visualize the similarity between items. Let us take a set of transfer and learning items placed on $\mathbb{R}^2$ (see Figure 3.6). The learning items are depicted as colored dots and are assigned to a specific category (A or B). The transfer items are depicted as black dots and are not members of a specific category (the location of the items is given by the center of the dots). To facilitate the comprehension, $\mathbb{R}^2$ is equipped with the Euclidean distance rather than the city-block distance. Let us take as reference point the transfer item in the middle (i.e., the black spot in the middle). The size of the spots in Figure 3.6 helps visualize the intensity of the similarity: the bigger the spots, the higher their similarity to the reference item (i.e., the item in the middle). By changing the value of the sensitive parameter c, the similarities between the reference point and the items of the space are altered. Thus, when $c = 0$, all exemplars are perceived as identical to the reference item. Conversely, when $c = 1$, only highly close items are perceived as similar to the reference item. ⊠

Step 3. We now have all the necessary elements to define the probability according to which the GCM classify items into a specific category. Let E_L be the set of learning items

and let $K_1, \ldots, K_N$ be the categories in which items are classified. According to the GCM, the probability of classifying item x into category $K \in \{K_1, \ldots, K_N\}$ (during the transfer phase) is given by:

$$\mathbb{P}(K \mid x) = \frac{\sum_{\xi \in K \cap E_L} S(\xi, x)}{\sum_{j=1}^{N} \sum_{\xi \in K_j \cap E_L} S(\xi, x)}. \tag{3.8}$$

We refer the reader to Box 3.2 for further information about the origin of Equation 3.8. In order to emphasize the parameter dependence of the probability, the probability of classifying a stimulus x into category K is denoted by $\mathbb{P}^{\theta}(K \mid x)$, where $\theta = (c, \omega_1, \ldots, \omega_{\mathfrak{N}})$ is the set of parameters of the model (i.e., the sensitive parameter c and the attention-weight parameters ω_i, for $i = 1, \ldots, \mathfrak{N}$). We underline that item x can be a learning or a transfer item.

Remark 3.4. Equation 3.8 allows us to understand why the GCM is classified as a transfer model. Indeed, the probability of classifying an item into a specific category is only determined by the set of learning items and does not evolve during time. Therefore, the GCM is only suitable for reproducing transfer performance. Moreover, the hypothesis according to which transfer performance does not evolve during the transfer phase seems appropriate since *i)* no feedback is provided during this phase, and *ii)* the transfer phase is usually short. ⊠

The intuitive idea behind the GCM is the following: if a new item is similar to the stored members of category K and less similar to the stored members of the other categories, then a learner would tend to classify it into category K. Thus, the higher the sum of similarities between the new item and the (learning) exemplars of category K (as compared to the contrasting categories), the higher the probability of classifying it into category K. The mechanism underlying the GCM is not based on abstraction processes and is highly costly.

Example 3.5. This example aims to extend Example 3.4 and showing how the probabilities of the GCM are influenced by different values of c. Our to-be-classified item is the point located in the middle (see Figure 3.6). The evidence in favor of category A and B are compared to determine the classification of the to-be-classified item. The evidence in favor of category A is given by the sum of similarities between the to-be-classified item and the exemplars of category A. The evidence in favor of category A can be visualized overlaying the turquoise blue points (the height of the obtained column reflects the category A evidence). Once the category A evidence found, it has to be normalized with respect to the sum of all evidences. Figure 3.7 shows the normalized evidence in

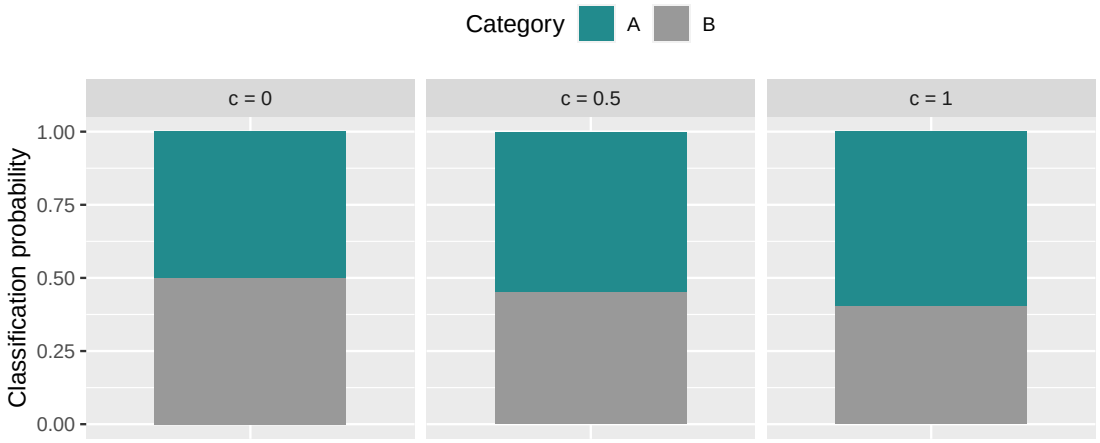


Figure 3.7 – *Normalized evidence, respect to the to-be-classified item of Figure 3.6 (i.e., the item in the middle), in favor of categories A and B as a function of the sensitive parameter c*

favor of categories A and B, depending on the value of c. When $c = 0$, the probability of classifying the to-be-classified item into one of the two categories are identical (i.e., $\frac{1}{2}$). Conversely, when $c = 1$, the probability of classifying the to-be-classified item into category A is higher as compared to category B. ⊠

Remark 3.5. Equation 3.8 is the version of the GCM that we considered. However, a more general version is available. In the most general version of the GCM Equation 3.8 is replaced with the following expression:

$$\mathbb{P}^{\theta}(K \mid x) = \frac{b_A \Big[\sum_{\xi \in K \cap E_L} V_{\xi K}\, S(\xi, x)\Big]^{\gamma}}{\sum_{j=1}^{N} b_{K_j} \Big[\sum_{\xi \in K_j \cap E_L} V_{\xi K_j}\, S(\xi, x)\Big]^{\gamma}}, \tag{3.9}$$

where $b_{K_j} \geq 0$ denotes the response-bias parameter of category K_j; $V_{\xi K_j}$ denotes the memory strength of item ξ with respect to category K_j; γ is a freely estimated response-scaling parameter; and $\theta = (c, \omega_1, \ldots, \omega_{\mathfrak{N}}, \gamma, b_{K_1}, \ldots, b_{K_N})$ represents the set of parameters of the model.

The memory-strength values $V_{\xi K_j}$ are generally not free parameters. The value $V_{\xi K_j}$ is typically set equal to the relative frequency with which item ξ is provided with category K_j feedback during the learning phase. Since exemplars are generally presented with equal frequency and assigned to only one category, this means that $V_{\xi K_j}$ is usually set equal to 1 when ξ belongs to category K_j and 0 otherwise. The parameter γ regulates the strength the model gives to the evidence in favor of the categories. ⊠

Computational Aspects

The aim of this paragraph is to describe the way the GCM was coded. As previously seen, the parameters of the GCM are the sensitive parameter c and the attention-weight parameters $\omega_1, \ldots, \omega_{\mathfrak{N}}$ ($i = 1, \ldots, \mathfrak{N}$) with the following conditions:

$$c \geq 0, \qquad \omega_i \geq 0, \qquad \sum_{i=1}^{\mathfrak{N}} \omega_i = 1.$$

However, we preferred to consider the equivalent following parameters:

$$\alpha_i = c \cdot \omega_i,$$

with $\alpha_i \geq 0$ (for all $i = 1, \ldots, \mathfrak{N}$) to avoid to deal with the last listed constraint. The sensitive and attention-weight parameters can be easily recovered using the following formulas:

$$c = \sum_{i=1}^{\mathfrak{N}} \alpha_i, \qquad \omega_i = \frac{\alpha_i}{\sum_{i=1}^{\mathfrak{N}} \alpha_i},$$

where $i = 1, \ldots, \mathfrak{N}$.

BOX 3.2 *Relation between Identification and Classification*

The context model and the GCM share the same formula for computing the probability of classifying stimuli into a specific category (see Equation 3.4 and 3.8). The aim of this box is to shed light on the origin of this equation. In [Nos84], Nosofsky has noticed a strong similarity between the formula mentioned above (Equation 3.4) and the Similarity-Choice Model (SCM) [Luc63; She57].

The similarity-choice model was introduced by Luce in 1963 to predict participant's performance in identification tasks. In an identification task, the participant is requested to assign a unique response to each of the stimuli (it can be considered as a classification task in which there are as many categories as the number of stimuli).

Nosofsky has found that the combination of the Luce's similarity-choice model with the mapping hypothesis led to Equation 3.4 (at least from a theoretical point of view). Let us explain this affirmation.

Let us consider an identification task where there are n stimuli to identify. According to the similarity-choice model, the probability to identify stimulus i with stimulus j is given by:

$$\mathbb{P}(j \mid i) = \frac{b_j s_{ij}}{\sum_{k=1}^{n} b_k s_{ik}}, \tag{3.10}$$

where b_j is the bias for response j and s_{ij} is the similarity between stimuli i and j. Even at this stage, the bias-free version of Equation 3.10 has a strike resemblance to Equation 3.4. The element that enables to connect the choice model with the context model is the mapping hypothesis. This hypothesis states that the probability of classifying stimulus i as a member of category K can be obtain by summing, over the category K stimuli, the probabilities to identify stimulus i with a stimulus of category K. The mathematical formalization of the mapping hypothesis is given by:

$$\mathbb{P}(K \mid i) = \sum_{j \in K} \mathbb{P}(j \mid i), \tag{3.11}$$

where $\mathbb{P}(K \mid i)$ is the probability of classifying stimulus i into category K and $\mathbb{P}(j \mid i)$ is the probability to identify stimulus i as stimulus j. According to the mapping hypothesis, classification performance can be predicted from identification performance.

Combining the Luce's similarity-choice model with the mapping hypothesis, the context model can be recovered. Indeed, by applying the mapping hypothesis to the bias-free version of Luce's similarity-choice model, we obtain that:

$$\mathbb{P}(K \mid i) = \sum_{j \in K} \mathbb{P}(j \mid i) = \frac{\sum_{j \in K} s_{ij}}{\sum_{k=1}^{n} s_{ik}}. \tag{3.12}$$

Equation 3.12 is identical to Equation 3.4, if we denote by E_L the set of n stimuli of the identification task and we assume that $K \subseteq E_L$. Although from a theoretical framework classification performance could be predicted from identification performance, this argument has been rejected on empirical grounds. Shepard et al. [SHJ61; She64] has observed systematic failure in predicting classification performance from identification performance. However, Nosofsky [Nos84] has proposed to keep this hypothesis, arguing that the key idea to preserve the mapping hypothesis is to assume that the similarity in identification and classification tasks are not equivalent.

TO SUM UP *Generalized Context Model (GCM)*

The Generalized Context Model (GCM) [MS78; Nos84; Nos86] is one of the most famous exemplar models in categorization. According to the GCM, the probability of classifying item x into category $K \in \{K_1, \ldots, K_N\}$ during the transfer phase is given by:

$$\mathbb{P}^{\theta}(K \mid x) = \frac{\sum_{\xi \in K \cap E_L} S(\xi, x)}{\sum_{j=1}^{N} \sum_{\xi \in K_j \cap E_L} S(\xi, x)}, \tag{3.13}$$

where

$$S(\xi, x) = e^{-cd(\xi,x)^p}, \qquad d(\xi, x) = \left[\sum_{i=1}^{\mathfrak{N}} \omega_i \cdot |\xi_i - x_i|^r\right]^{\frac{1}{r}},$$

N is the number of categories, E_L is the set of learning items, $\mathfrak{N}$ is the dimension of the psychological space in which exemplars are embedded, c is a sensitive parameter, $\omega_1, \ldots, \omega_{\mathfrak{N}}$ are attention-weight parameters, and p and r are positive constants. The parameters of the GCM are:

$$c \geq 0 \qquad \text{and} \qquad \omega_1, \ldots, \omega_{\mathfrak{N}} \geq 0,$$

with the condition $\sum_{i=1}^{\mathfrak{N}} \omega_i = 1$.

3.2.2 Ordinal General Context Model (OGCM)

The Generalized Context Model is not sensitive to different presentation orders. Indeed, according to the GCM, the probability of classifying new items into a specific category is only related to physical features of the set of learning items. However, different presentation orders can shape the way we perceive, represent and learn information (see Chapter 1 Section 1.3). Here, we propose a modification of the GCM that accounts for the order in which stimuli are presented.

This new exemplar model is called Ordinal General Context Model (OGCM) and integrates in the GCM's structure a component aiming at capturing ordinal effects. We declined the OGCM into three versions, each of them integrating a different ordinal aspect: *i)* the OGCM-L incorporates the average presentation order received during the learning phase, *ii)* the OGCM-M incorporates the most frequent (i.e., the median) presentation order

received during the learning phase, and *iii)* the OGCM-T incorporates the presentation order received during the transfer phase.

Mathematical Description

As for the GCM, the stimuli are represented as points in a multidimensional psychological space. However, instead of considering a distance that is exclusively related to physical features, the OGCM adopts a distance that integrates ordinal aspects. More specifically, the distance considered by the OGCM is defined as follows:

$$\mathcal{D}(\xi, x) = \left[\sum_{i=1}^{\mathfrak{N}} \omega_i \cdot |\xi_i - x_i|^r + \omega_O \cdot d_O(\xi, x)^r \right]^{\frac{1}{r}}, \tag{3.14}$$

where the first term is the feature-distance defined in the GCM (see Equation 3.6), $d_O(\xi, x)$ is the ordinal distance between items ξ and x, and ω_O is the attention-weight related to the ordinal dimension. Since both Experiment I and II involved highly separable-dimension stimuli, the value of r is set equal to 1.

Equation 3.14 is the only equation differentiating the OGCM from the GCM. Both the similarity between two items and the probability of classifying an item into a specific category are the same as in the GCM (Equation 3.7 and 3.8). Before defining the ordinal distance $d_O(\xi, x)$ for each of the versions of the OGCM, let us consider an example to understand the concept of ordinal distance within a block.

Example 3.6. Let us consider four items (a, b, c, and d) placed in a two-dimensional psychological space (see Figure 3.8, on the left). Let us assume that these four items are presented to a participant in a specific order. The order manipulation is shown in Figure 3.8 (in the middle) by means of numbers: item a was presented first, item d second, and so on. The order manipulation can be visualize adding a dimension in which item a takes value 1, item d takes value 2, and so on (see Figure 3.8, on the right).

The ordinal distance between two items is defined as the difference of their ordinal position in the order manipulation (the difference of values in the additional dimension). For instance, the ordinal distance between items a and d is 1, the one between items d and b is two, and so on. ⊠

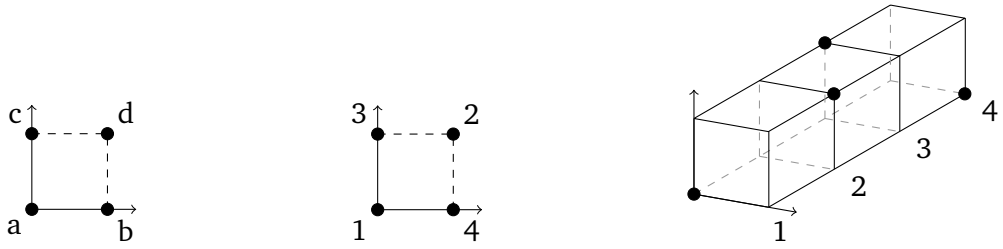


Figure 3.8 – *Visualization of the ordinal distance within a block. On the left, the spacial position of the items (a, b, c, and d). On the middle, the presentation order in which stimuli were presented. On the right, the addition of a dimension accounting for the order manipulation.*

Each version of the OGCM defines the ordinal distance $d_O(\xi, x)$ in different ways.

OGCM-L. In the OGCM-L, the ordinal distance d_O between two items averages across the learning phase the ordinal distances within a block between the two considered items (if they are both learning items). Conversely, their ordinal distance is set equal to the maximal distance among the ordinal distances plus 1.

OGCM-M. The OGCM-M is similar to the OGCM-L. However, instead of averaging the ordinal distances, the OGCM-M considers the median of the ordinal distances. Therefore, the ordinal distance d_O between two learning items is set equal to the median across the learning phase of their ordinal distances within a block. Conversely, their ordinal distance is set equal to the maximal distance among the ordinal distances plus 1.

OGCM-T. In the OGCM-T, the ordinal distance d_O between two items is defined as the ordinal distance between the two items within the transfer block.

Remark 3.6. The ordinal distance d_O between a transfer item and a learning item do not have to be set at the maximal distance among the ordinal distances plus 1. The rational was to set a value greater than the maximal ordinal distance between two learning items. Indeed, one intuitive hypothesis is that, in the ordinal dimension, the psychological distance between a new item and a stored item is perceived as higher than the psychological distance between two stored items. Moreover, the ordinal distance can be re-scaled to ensure that both feature and ordinal dimensions share the same order of magnitude. ⊠

Example 3.7. Let us give an example about how the three versions of the OGCM define the ordinal distance d_O between two items. Let us consider three learning items (a gray

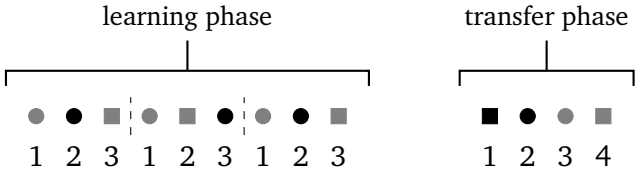


Figure 3.9 – *Visualization of Example 3.7. The learning items are a gray circle, a black circle, and a gray square. The transfer item is a black square.*

circle, a black circle and a gray square) and one transfer item (a black square). The three learning items are presented during the learning phase, while the full set of learning and transfer items is presented during the transfer phase. Let us assume that the learning phase is composed of three blocks, while the transfer phase is composed of one block (see Figure 3.9).

According to the OGCM-L, the distance between the gray and black circles is obtained averaging their ordinal position across the learning phase (i.e., $\frac{1+2+1}{3}$). Conversely, the distance between the gray circle and black square is set equal to a constant (in this case we set it equal to 3, i.e. the maximal ordinal distance in the learning phase plus 1).

According to the OGCM-M, the distance between the gray and black circles is obtained taking the median of their ordinal position across the learning phase (i.e., 1). Conversely, the distance between the gray circle and black square is set equal to a constant (in this case we set it equal to 3, i.e. the maximal ordinal distance in the learning phase plus 1).

According to the OGCM-T, the distance between the gray and black circles is obtained taking their ordinal position within the transfer block (i.e., 1). The same goes for the distance between the gray circle and black square which is 2. ⊠

Computational Aspects

Similarly to the GCM, instead of considering the parameters c and ω_i ($i = 1, \ldots, \mathfrak{N}$) with the constraints $c \geq 0$, $\omega_i \geq 0$, and $\sum_{i=1}^{\mathfrak{N}} \omega_i = 1$, the following equivalent parameters $\alpha_i = c \cdot \omega_i$ (with the constrains $\alpha_i > 0$) were preferred.

TO SUM UP *Ordinal General Context Model (OGCM)*

The Ordinal General Context Model (OGCM) is a modification of the GCM that accounts for the order in which stimuli are presented. The OGCM was declined in three versions:

OGCM-L. This version integrates the average presentation order that participants received during the learning phase.

OGCM-M. This version integrates the most frequent presentation order that participants received during the learning phase.

OGCM-T. This version integrates the average presentation order that participants received during the transfer phase.

According to the versions of the OGCM, the probability of classifying item x into category $K \in \{K_1, \ldots, K_N\}$ (during the transfer phase) is given by:

$$\mathbb{P}^{\theta}(K \mid x) = \frac{\sum_{\xi \in K \cap E_L} S(\xi, x)}{\sum_{j=1}^{N} \sum_{\xi \in K_j \cap E_L} S(\xi, x)}, \tag{3.15}$$

where

$$S(\xi, x) = e^{-c\mathcal{D}(\xi,x)^p}, \qquad \mathcal{D}(\xi, x) = \left[\sum_{i=1}^{\mathfrak{N}} \omega_i \cdot |\xi_i - x_i|^r + \omega_O \cdot d_O(\xi, x)^r \right]^{\frac{1}{r}},$$

N is the number of categories in which stimuli can be classified, E_L denotes the set of learning items, c is a sensitive parameter, p and r are positive constants, $\omega_1, \ldots, \omega_{\mathfrak{N}}$ are attention-weight parameters, ω_O is the attention-weight parameter related to the ordinal dimension, and $d_O(\xi, x)$ is the ordinal distance between items ξ and x. The parameters of the model are:

$$c \geq 0 \qquad \text{and} \qquad \omega_1, \ldots, \omega_{\mathfrak{N}}, \omega_O \geq 0,$$

with the condition $\sum_{i=1}^{\mathfrak{N}} \omega_i + \omega_O = 1$.

3.2.3 Generalized Context Model equipped with the Lag Mechanism (GCM-Lag)

Again, the Generalized Context Model is not sensitive to different presentation orders. However, a sequence-sensitive version of the GCM has been developed to palliate this disadvantage [NKM92]. This sequence-sensitive version of the GCM is based on the following principle: rather that giving equal weight to all items in the computation of the classification probability, greater weight is given to more recently presented items. In other words, a lag mechanism that accounts for sequential effects has been integrated to the GCM. Our aim was to complete the picture by including this sequence-sensitive version of the GCM.

Mathematical Description

Similarly to the GCM, the stimuli are represented as points in a multidimensional psychological space and the distance between two items is computed as in Equation 3.6. The similarity between two items is also computed as in the GCM (Equation 3.7). However, the classification probability is defined differently. According to the GCM-Lag, the probability of classifying item x into category $K \in \{K_1, \ldots, K_N\}$ (during the transfer phase) is given by:

$$\mathbb{P}(K \mid x) = \frac{\sum_{\xi \in K \cap E_L} \mathcal{R}_{\xi,x} \cdot S(\xi, x)}{\sum_{j=1}^{N} \sum_{\xi \in K_j \cap E_L} \mathcal{R}_{\xi,x} \cdot S(\xi, x)}, \tag{3.16}$$

where $\mathcal{R}_{\xi,x}$ is the memory strength associated with exemplar ξ. The memory strength is defined as an exponential decay function of the lag presentation as follows:

$$\mathcal{R}_{\xi,x} = e^{-\delta \cdot \text{lag}(\xi,x)}.$$

The quantity $\text{lag}(\xi, x)$ represents the number of intervening trials between the presentations of stimuli ξ and x, while δ is a time-rate decay parameter.

Computational Aspects

Similarly to the GCM, instead of considering the parameters c and ω_i ($i = 1, \ldots, \mathfrak{N}$) with the constraints $c \geq 0$, $\omega_i \geq 0$, and $\sum_{i=1}^{\mathfrak{N}} \omega_i = 1$, the following equivalent parameters $\alpha_i = c \cdot \omega_i$ (with the constrains $\alpha_i > 0$) were preferred.

TO SUM UP *GCM Equipped with the Lag Mechanism (GCM-Lag)*

The GCM-Lag is a sequence-sensitive version of the GCM [NKM92]. According to the GCM-Lag, the probability of classifying item x into category $K \in \{K_1, \ldots, K_N\}$ during the transfer phase is given by:

$$\mathbb{P}^{\theta}(K \mid x) = \frac{\sum_{\xi \in K \cap E_L} \mathcal{R}_{\xi,x} \cdot S(\xi, x)}{\sum_{j=1}^{N} \sum_{\xi \in K_j \cap E_L} \mathcal{R}_{\xi,x} \cdot S(\xi, x)}, \tag{3.17}$$

where

$$S(\xi, x) = e^{-cd(\xi,x)^p}, \qquad \mathcal{R}_{\xi,x} = e^{-\delta \cdot \text{lag}(\xi,x)}, \qquad d(\xi, x) = \left[\sum_{i=1}^{\mathfrak{N}} \omega_i \cdot |\xi_i - x_i|^r\right]^{\frac{1}{r}},$$

N is the number of categories, E_L is the set of learning items, $\mathfrak{N}$ is the dimension of the psychological space in which exemplars are embedded, c is a sensitive parameter, $\omega_1, \ldots, \omega_{\mathfrak{N}}$ are attention-weight parameters, δ is a time-rate decay parameter, $\text{lag}(\xi, x)$ is the number of intervening trials between the presentations of stimuli ξ and x, and p and r are positive constants. The parameters of the GCM-Lag are:

$$c \geq 0, \qquad \delta \geq 0 \qquad \text{and} \qquad \omega_1, \ldots, \omega_{\mathfrak{N}} \geq 0,$$

with the condition $\sum_{i=1}^{\mathfrak{N}} \omega_i = 1$.

3.2.4 Relations Between Transfer Models

This subsection aims to show the relations and inclusions between the selected transfer models.

Proposition 3.1. *The Generalized Context Model (GCM) is included in both the three versions of the Ordinal General Context Model (OGCM) and the GCM equipped with the lag mechanism (GCM-Lag).*

Proof. The GCM can be obtained from the three versions of the OGCM by setting the parameter related to the ordinal dimension (i.e., ω_O) equal to 0. It can also be obtained from the GCM-Lag by setting the time-rate decay parameter (i.e., δ) equal to 0. □

Since the context model was described in Box 3.1 and it is strictly related to the GCM, a mathematical comparison between the GCM and the context model is also given (see Box 3.3).

BOX 3.3 *Comparison Between the GCM and the Context Model*

Proposition 3.2. *The Generalized Context Model (GCM) is included in the context model when the dimensions of the stimuli are binary-valued/Boolean and* $p = r$.

Proof. If $p = r$, we have that:

$$S(\xi, x) = e^{-c \cdot \sum_{i=1}^{\mathfrak{N}} \omega_i \cdot |\xi_i - x_i|^r} = \prod_{i=1}^{\mathfrak{N}} e^{-c \cdot \omega_i \cdot |\xi_i - x_i|^r}.$$

If there are only two options for each dimension (let us say 0 and 1), we have that:

$$|\xi_i - x_i| = \begin{cases} 1 & \text{if } \xi_i \neq x_i \\ 0 & \text{if } \xi_i = x_i \end{cases}$$

Therefore $|\xi_i - x_i|^r$ is equal to the $\delta_i(\xi, x)$ defined in the context model:

$$|\xi_i - x_i|^r = \begin{cases} 1 & \text{if } \xi_i \neq x_i \\ 0 & \text{if } \xi_i = x_i \end{cases} = \delta_i(\xi, x)$$

The similarity between two items can thus be written as:

$$S(\xi, x) = \prod_{i=1}^{\mathfrak{N}} e^{-c \cdot \omega_i \cdot |\xi_i - x_i|^r} = \prod_{i=1}^{\mathfrak{N}} e^{-c \cdot \omega_i \cdot \delta_i(\xi, x)} = \prod_{i=1}^{\mathfrak{N}} \left[e^{-c \cdot \omega_i}\right]^{\delta_i(\xi, x)}.$$

If we define $s_i = e^{-c \cdot \omega_i}$, then we have that $S(\xi, x) = \prod_{i=1}^{\mathfrak{N}} s_i^{\delta_i(\xi, x)}$ □

3.3 Learning Models

The aim of this section is to describe the learning models that were used to investigate both category learning and transfer. Indeed, learning models are capable to reproduce both the learning dynamics and transfer performance. The selected learning models are the following: the Component-Cue model and ALCOVE.

The rationale was to compare two models with a similar mathematical structure but implementing two different strategies. Indeed, although both models are based on artificial neural networks (see Box 3.4), they actualize either a rule-based (the Component-Cue model) or a similarity-based strategy (ALCOVE). A limited number of studies have investigated the performance of both models or the one of their extensions [GB88a; GB88b; GBH89; Kru92; NKM92; Nos+94; Pal99]. Our aim is to further investigate these models to determine the one that best fits our data as well as to search for a relation between presentation order and learning strategies.

Box 3.4 *Artificial Neural Network*

Artificial Neural Network (ANN) [Dre90; Ros58; WJ74] (also called connectionist systems) are systems that are able to learn how to perform supervised tasks through trial and error. For example, an artificial neural network can be trained to identify dogs by providing images that have been manually labeled “dog” or “no dog”.

The basic unit of an ANN is a node, also called a neuron (see Figure 3.10). Artificial neural networks are composed of three main layers: a layer of input nodes that receive the information; one or several layers of intermediate nodes (called hidden nodes) that elaborate the information; and a layer of output nodes that represent the outputs of the network.

The nodes of a layer (except the output nodes) are connected to the nodes of the following layer by means of weights. All weights are initiated at a specific value and evolve every time a new input is received. Signals travel from the input layer to the output layer, passing through the hidden layer (or layers). When the signal reaches the output layer, the difference between the response of the network and the

feedback is computed. The process consisting in the reception of new information, its elaboration, and the computation of the error is called forward propagation.

The forward propagation is followed by a backward propagation. During the backward propagation, the weights on the connections (between nodes) are updated to minimize the error of the network. The updating of the weights is generally implemented via gradient descent.

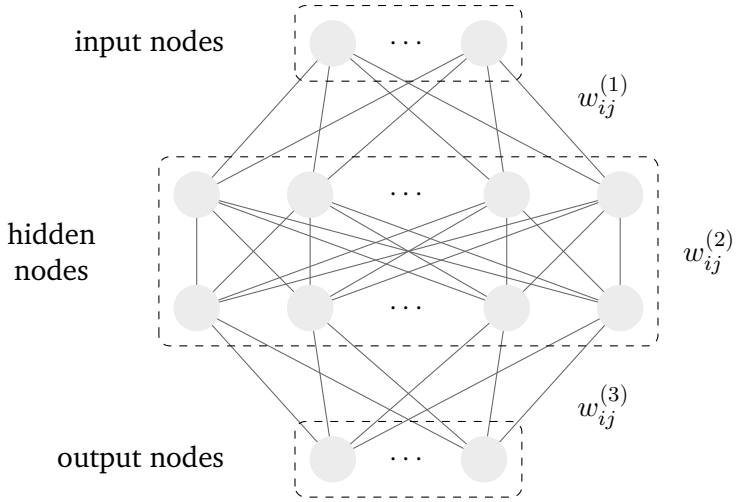


Figure 3.10 – *Structure of a general artificial neural network.*

3.3.1 Component-Cue Model

The Component-Cue model was introduced by Gluck and Bower in 1988 [GB88b; GBH89]. Originally labeled as adaptive model, nowadays it is labeled as artificial neural network.

Mathematical Description

The Component-Cue model is based on an artificial neural network. Such a structure (see Figure 3.11) is composed of of three layers: a node receiving the stimuli (input node), a layer of intermediate nodes that elaborate the new information (feature nodes), and a layer of output nodes that generate an output for each category (category nodes). Feature nodes are linked to category nodes by means of weights. The evolution of these

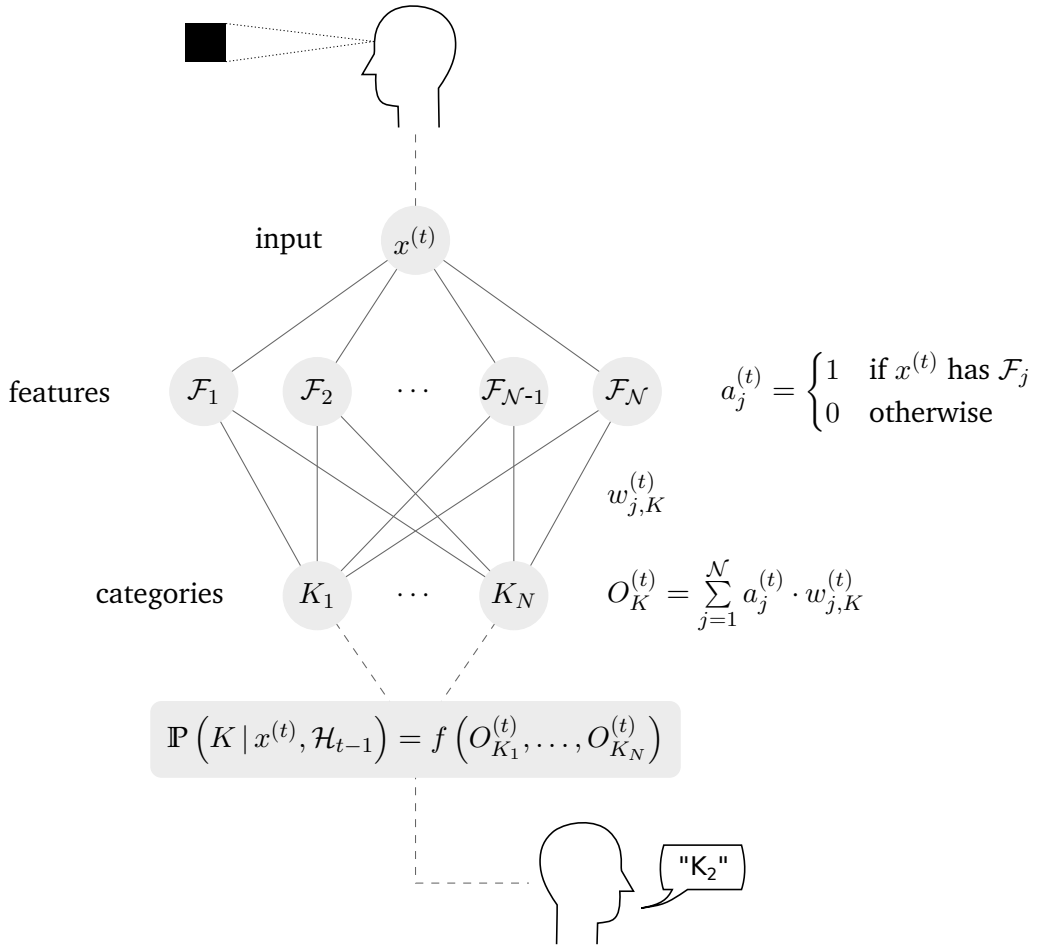


Figure 3.11 – *Artificial neural network structure of Component-Cue.*

weights allows the model to learn. After this brief presentation, let us describe the Component-Cue model in detail.

When a stimulus of the learning phase is presented to a participant, it is received by the input node. The signal is then sent to the feature nodes. Each feature node codes a particular feature that exemplars have. For instance, if the stimuli are a gray circle and a black circle, then the features discriminating the two items are “gray” and “black”. Therefore, the network has two feature nodes, one coding the feature “gray” and the other coding the feature “black”. Feature nodes are activated depending whether the input stimulus is characterized by the feature coded by the node. For instance, in the previous example, when a gray stimulus is presented, the feature node “gray” is activated with a value of 1 and the feature node “black” is turned off (or, equivalently, it is activated

with a value of 0). Conversely, when a black stimulus is presented, the feature node "gray" is turned off and the feature node "black" is activated with a value of 1.

More precisely, let us assume that the items presented during the classification task (the learning phase specifically) have $\mathcal{N}$ features $\mathcal{F}_1, \ldots, \mathcal{F}_{\mathcal{N}}$. Let $x^{(t)}$ be the t-th stimulus presented to the participant. Each feature node is activated by the quantity:

$$a_j^{(t)} = \begin{cases} 1 & \text{if } x^{(t)} \text{ has } \mathcal{F}_j \\ 0 & \text{otherwise} \end{cases}$$

where $a_j^{(t)}$ ($j = 1, \ldots, \mathcal{N}$) represents the activation of the j-th feature node due to the presentation of the t-th stimulus.

Remark 3.7. The dimensions (i.e., the features) of the stimuli of Experiment I and II are binary-valued. Therefore, the number of feature nodes of the Component-Cue model is $\mathcal{N} = 2 \times \mathfrak{N}$, where $\mathfrak{N}$ represents the dimension of the psychological space in which stimuli are embedded. ⊠

Once the feature nodes are activated, the signal is sent to the output nodes. Each output node codes a category in which items are classified. The activation of the feature node is multiplied by a weight (the weight that links feature nodes and categories) and these weighted activations are then summed to form outputs. More precisely, let us suppose that there are N categories $K_1, \ldots, K_N$. The output node $K \in \{K_1, \ldots, K_N\}$ is activated by the quantity:

$$O_K^{(t)} = \sum_{j=1}^{\mathcal{N}} a_j^{(t)} \cdot w_{j,K}^{(t)},$$

as a result of the reception of the t-th stimulus, where $w_{j,K}^{(t)}$ is the weight linking feature node $\mathcal{F}_j$ to category node K. The weight $w_{j,K}^{(t)}$ ($j = 1, \ldots, \mathcal{N}$) are called the association weights.

Once the outputs of the categories are computed, the model computes the probability of classifying an item into a specific category. This probability is a function of the category outputs. If the model is asked, it classifies stimuli on the basis of the computed probabilities. Two different formulas to compute the classification probability has been used in the literature: an exponential formula and a linear formula.

Exponential version. According to the exponential version, the probability of classifying the t-th stimulus $x^{(t)}$ into category $K \in \{K_1, \ldots, K_N\}$ (knowing the history of the process $\mathcal{H}_{t-1}$) is given by:

$$\mathbb{P}(K \mid x^{(t)}, \mathcal{H}_{t-1}) = \frac{e^{\phi O_K^{(t)}}}{\sum_{j=1}^{N} e^{\phi O_{K_j}^{(t)}}}, \tag{3.18}$$

where ϕ is a freely estimated positive parameter. This version appeared in the Gluck and Bower's paper [GB88b]. The exponential version is denoted by an E (i.e., Component-Cue$^{\mathrm{E}}$).

Linear version. According to the linear version, the probability of classifying the t-th stimulus $x^{(t)}$ into category $K \in \{K_1, \ldots, K_N\}$ (knowing the history of the process $\mathcal{H}_{t-1}$) is given by:

$$\mathbb{P}(K \mid x^{(t)}, \mathcal{H}_{t-1}) = \frac{O_K^{(t)} + b}{\sum_{j=1}^{N} \left(O_{K_j}^{(t)} + b\right)} \tag{3.19}$$

where b is a category bias parameter. To our knowledge, this version has not been applied to the Component-Cue model. However, it is often used for ALCOVE. Therefore, both versions were considered to provide a complete comparison between the two models. The linear version is denoted by an L (i.e., Component-Cue$^{\mathrm{L}}$).

Remark 3.8. The exponential and linear versions are two different ways to constraint the outputs to be positive and smaller then 1. However, the parameters of the versions have two distinct psychological interpretations. In the exponential version, the higher the parameter (i.e., ϕ), the greater the influence of the higher output. For instance, if ϕ is higher than 1 and O_K is the higher output as a result of the presentation of a stimulus, the perception that the stimulus is closer to category K than other categories is amplified. Conversely, in the linear version, the higher the parameter, the smaller the influence of the higher output. For instance, if b is higher than 1 and O_K is the higher output as a result of the presentation of a stimulus, the perception that the stimulus is closer to category K than other categories is weakened. ⊠

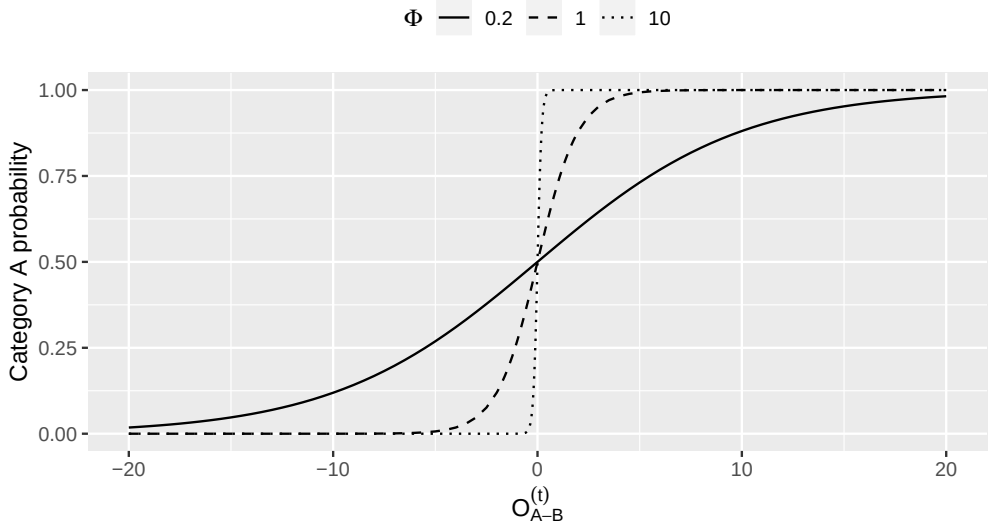


Figure 3.12 – *Exponential classification probability of the Component-Cue model as a function of the parameter ϕ when only two categories (A and B) are considered. On the x-axis, the difference between the two outputs. On the y-axis, the probability of classifying a stimulus as a member of category A.*

Remark 3.9. When there are only two categories (A and B), Equation 3.18 can be expressed as a function of a single variable. Let us set $O_{A-B}^{(t)} = O_A^{(t)} - O_B^{(t)}$ and let us replace $O_B^{(t)}$ with $O_A^{(t)} - O_{A-B}^{(t)}$ in Equation 3.18, then we have that:

$$\begin{aligned}\mathbb{P}(A \,|\, x^{(t)}, \mathcal{H}_{t-1}) &= \frac{e^{\phi O_A^{(t)}}}{e^{\phi O_A^{(t)}} + e^{\phi O_B^{(t)}}} \\ &= \frac{e^{\phi O_A^{(t)}}}{e^{\phi O_A^{(t)}} + e^{\phi O_A^{(t)}} \cdot e^{-\phi O_{A-B}^{(t)}}} \\ &= \frac{1}{1 + e^{-\phi O_{A-B}^{(t)}}}. \end{aligned} \tag{3.20}$$

Figure 3.12 shows the exponential classification probability as a function of the parameter ϕ when only two categories are considered. ⊠

Once the classification probability of the input stimulus is computed, the association weights $w_{j,K}^{(t)}$ are updated. The evolution of the weights allows the model to learn. The updating of the association weights is based on the gradient descent algorithm, whose aim is to minimize (trial to trial) the error between the output of the model and the

provided feedback. More precisely, the error generated by the model after the reception of the t-th stimulus is computed as follows:

$$E^{(t)} = \sum_{j=1}^{N} \left(\mathcal{T}_{K_j}^{(t)} - O_{K_j}^{(t)} \right)^2, \tag{3.21}$$

where

$$\mathcal{T}_{K_j}^{(t)} = \begin{cases} 1 & \text{if } x^{(t)} \in K_j \\ -1 & \text{otherwise} \end{cases} \tag{3.22}$$

Equation 3.22 represents the feedback given to the model (which is a numerical equivalent of the feedback given to the participant). For instance, if the t-th stimulus belongs to category $K \in \{K_1, \ldots, K_N\}$, the feedback given to the model is that the category K output should have been activated by the quantity 1, while the other outputs should have been activated by the quantity -1.

The association weights are updated as follows to decrease the error of the model:

$$\begin{aligned} w_{j,K}^{(t+1)} &= w_{j,K}^{(t)} - \lambda_w \cdot \frac{\partial E^{(t)}}{\partial w_{j,K}^{(t)}} \\ &= w_{j,K}^{(t)} + \lambda_w \cdot a_j^{(t)} \cdot \left(\mathcal{T}_K^{(t)} - O_K^{(t)} \right), \end{aligned} \tag{3.23}$$

where λ_w is a freely learning rate parameter, $j \in \{1, \ldots, \mathcal{N}\}$ and $K \in \{K_1, \ldots, K_N\}$. The association weights are initiated at 0:

$$w_{j,K}^{(0)} = 0,$$

for all $j \in \{1, \ldots, \mathcal{N}\}$ and $K \in \{K_1, \ldots, K_N\}$.

To recap, the parameters of the exponential version of the Component-Cue model are ϕ and λ_w, while those of the linear version are b and λ_w. The probability in Equation 3.18 and 3.19 is denoted by $\mathbb{P}^{\theta}(K \mid x^{(t)}, \mathcal{H}_{t-1})$ (where θ is the set of parameters of the model) to emphasize its dependency from the parameters.

The Component-Cue model (and learning model in general) can also be applied to reproduce transfer performance. Since feedback are not provided during the transfer phase, the updating of the weights of the network stops and a static variant of the model is considered. The "dinamic" version is applied to the learning phase, while the static version is applied to the transfer phase. In the static version, the association weights are

considered as parameters of the model, even though in a practical context the association weights are determined through the application of the "dinamic" version to the learning phase. The static variant of the model is denoted by Component-Cue-S^E (when the exponential version is considered) and Component-Cue-S^L (when the linear version is considered).

Remark 3.10. A more general version of the Component-Cue model is less often considered in the literature (see [NKM92]). In this more general version (that we do not implement), the activation of the output nodes is also expressed as a function of category-bias weights, connecting feature nodes to category nodes. More precisely, the activation of the output node $K \in \{K_1, \ldots, K_N\}$ as a result of the reception of the t-th stimulus is defined as follows:

$$O_K^{(t)} = \sum_{j=1}^{\mathcal{N}} a_j^{(t)} \cdot w_{j,K}^{(t)} + b_K^{(t)},$$

where $b_K^{(t)}$ is the category-bias weight linking feature node $\mathcal{F}_j$ to category node K. Similarly to the association weights, the category-bias weights are updated. The updating of these weights is given by the following rule:

$$\begin{aligned} b_K^{(t+1)} &= b_K^{(t)} - \lambda_b \cdot \frac{\partial E^{(t)}}{\partial b_K^{(t)}} \\ &= b_K^{(t)} + \lambda_b \cdot \left(\mathcal{T}_K^{(t)} - O_K^{(t)}\right), \end{aligned} \tag{3.24}$$

where λ_b is a freely learning rate parameter and $K \in \{K_1, \ldots, K_N\}$. ⊠

Remark 3.11. Gluck and Bower also developed a Configural-Cue model [GBH89] in which the input nodes code not only the individual feature of the stimuli but also the pairs of features, the triples of features, and so forth. ⊠

Example 3.8. Let us consider four exemplars (a black square, a gray square, a black circle and a gray circle) belonging to either category A or category B. The Component-Cue model applied to this case is illustrated in Figure 3.13. The intermediate nodes code the features of the items (square, circle, black, gray), while the output nodes code the categories in which the items are classified (A and B). ⊠

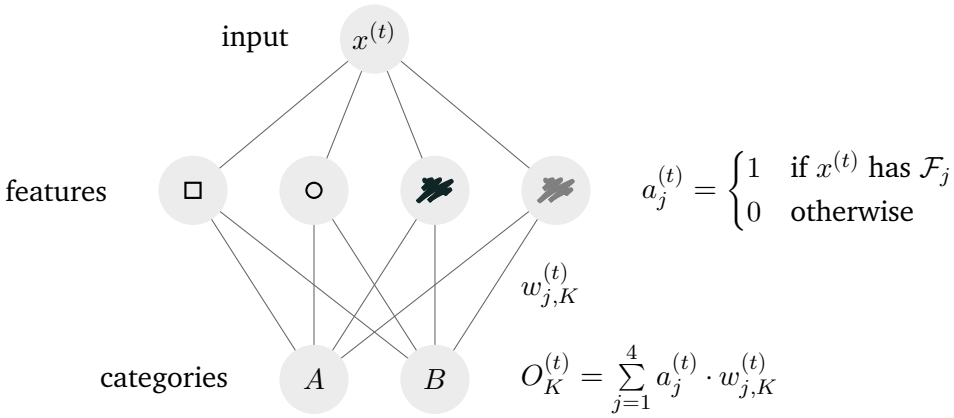


Figure 3.13 – *The Component-Cue model when four items (a black square, a gray square, a black circle and a gray circle) and two categories (A and B) are considered.*

Computational Aspects

The aim of this paragraph is to describe the way the Component-Cue model was coded. The following constraints were used:

i. The learning rate parameter of the association weights λ_w was constrained to be greater than 0.

ii. The outputs was constrained to be between 1 and -1. Therefore, when the outputs were either greater than 1 or smaller than -1, the association weights $w_{j,K}^{(t)}$ ($j \in \{1, \ldots, \mathcal{N}\}$ and $K \in \{K_1, \ldots, K_N\}$) were re-scaled as follows:

$$w_{j,K}^{(t)} \leftarrow \frac{w_{j,K}^{(t)}}{|O_K^{(t)}|}.$$

This condition ensure that the classification probability of the linear version was positive (Equation 3.19). However, to maintain coherence between the two versions, the same constraint was also used in the exponential version.

iii. The parameter b of the linear version was constrained to be grater than 1 to ensure that the classification probability in Equation 3.19 was positive.

TO SUM UP *Component-Cue Model*

The Component-Cue model is a learning model developed by Gluck and Bower [GB88a; GB88b], based on an artificial neural network structure. Its structure is composed of a single input node receiving the stimuli, a layer of feature nodes coding the features of the stimuli, and a layer of category nodes representing the categories in which stimuli are classified. When a stimulus $x^{(t)}$ reaches the input node, the feature nodes $\mathcal{F}_j$ $(j = 1, \ldots, \mathcal{N})$ are activated according to the quantity:

$$a_j^{(t)} = \begin{cases} 1 & \text{if } x^{(t)} \text{ has } \mathcal{F}_j \\ 0 & \text{otherwise} \end{cases}$$

where $a_j^{(t)}$ $(j = 1, \ldots, \mathcal{N})$ represents the activation of the j-th feature node due to the presentation of the t-th stimulus. The activations of the feature nodes are then multiplied by association weights, which are summed to form outputs. The K output node is activated by the quantity:

$$O_K^{(t)} = \sum_{j=1}^{\mathcal{N}} a_j^{(t)} \cdot w_{j,K}^{(t)},$$

where $w_{j,K}^{(t)}$ is the weight linking feature node $\mathcal{F}_j$ to category node K ($K = K_1, \ldots, K_N$).

Finally, the probability of classifying stimulus $x^{(t)}$ into category $K \in \{K_1, \ldots, K_N\}$ is given by:

$$\mathbb{P}\left(K \,|\, x^{(t)}, \mathcal{H}_{t-1}\right) = \frac{e^{\phi O_K^{(t)}}}{\sum_{j=1}^{N} e^{\phi O_{K_j}^{(t)}}} \qquad \text{or} \qquad \frac{O_K^{(t)} + b}{\sum_{j=1}^{N} \left(O_{K_j}^{(t)} + b\right)},$$

depending whether the exponential or linear version is considered. The quantities ϕ and b are two positive freely estimated parameters.

The association weights are updated using a gradient descent algorithm to minimize the error between the outputs of the model and the feedback. The updating of

the association weights is given by the following rule ($j \in \{1, \ldots, \mathcal{N}\}$ and $K \in \{K_1, \ldots, K_N\}$):

$$w_{j,K}^{(t+1)} = w_{j,K}^{(t)} + \lambda_w \cdot a_j^{(t)} \cdot \left(\mathcal{T}_K^{(t)} - O_K^{(t)}\right),$$

where λ_w is a freely learning rate parameter and

$$\mathcal{T}_K^{(t)} = \begin{cases} 1 & \text{if } x^{(t)} \in K \\ -1 & \text{otherwise.} \end{cases}$$

The association weights are initialized at 0. To recap, the parameters of the model are:

$$\lambda_w,\, \phi \qquad \text{or} \qquad \lambda_w,\, b$$

depending whether the exponential or linear version is considered. The Component-Cue model can also be applied to reproduce transfer performance stopping the updating of the association weights. We refer to this variant of the model as the static variant.

3.3.2 Attention Learning COVEring Map Model (ALCOVE)

The Kruschke's Attention Learning COVEring map model (ALCOVE) appeared for the first time in 1992 [Kru92]. It is a connectionist model (i.e. it is based on artificial neural network) and combines aspects of two models previously described, the GCM and Component-Cue.

On one hand, ALCOVE shares with the GCM an exemplar-based representation of the stimuli. Indeed, ALCOVE assumes that people store all exemplars they encounter and classify new stimuli on the basis of their similarity to these stored exemplars. Moreover, ALCOVE integrates an attention mechanism as the GCM.

On the other hand, ALCOVE shares with the Component-Cue model a trial and error dynamics. To recap, ALCOVE can be considered as the version of the GCM integrating an error-driven mechanism. According to Kruschke [Kru92], ALCOVE extends both the GCM, by including a learning mechanism, and Component-Cue, "by allowing continuous dimensions and including explicit dimensional attention learning".

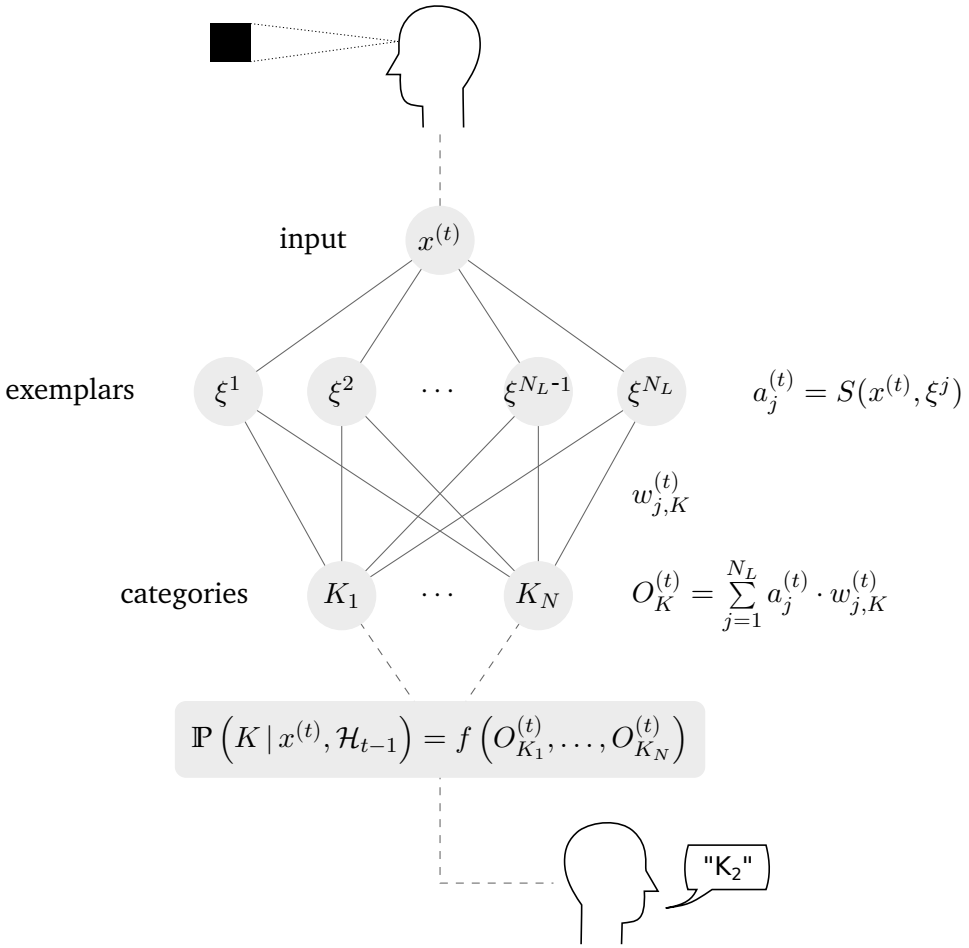


Figure 3.14 – *Artificial neural network structure of ALCOVE.*

Mathematical Description

The artificial neural network structure of ALCOVE is illustrated in Figure 3.14. Similarly to Component-Cue, ALCOVE is composed of three layers: a node receiving the stimuli (input node), a layer of intermediate nodes that elaborate the information (exemplar nodes), and a layer of output nodes that generate an output for the categories in which stimuli are classified (category nodes). Exemplar nodes are linked to category nodes by means of weights. Again, the evolution of these weights allows the model to learn. After this brief presentation, let us describe ALCOVE in detail.

When a stimulus is presented, it is received by the input node. The reception of the stimulus produces the activation of the exemplar nodes. The exemplar nodes code the learning items of the classification task and they are activated according to their similarity to the input stimulus. The similarity between two items is defined as in the GCM. Therefore, as a result of the reception of stimulus $x^{(t)}$, the exemplar node ξ^j (in $\xi^1, \ldots, \xi^{N_L}$) is activated by the following quantity:

$$a_j^{(t)} = S(x^{(t)}, \xi^j) = e^{-c \cdot d(x^{(t)}, \xi^j)^p}, \tag{3.25}$$

where $S(x^{(t)}, \xi^j)$ denotes the similarity between exemplars $x^{(t)}$ and ξ^j; c is a freely estimated sensitive parameter; p is a positive constant determined on the basis of the nature of the stimuli (see Subsection 3.2.1 for further information about the choice of p); and $d(x^{(t)}, \xi^j)$ represents the distance between exemplars $x^{(t)}$ and ξ^j. Since both Experiment I and II involved distinguishable stimuli, the value of p is set equal to 1.

The distance between two items is also computed as in the GCM. Items are considered as points in a $\mathfrak{N}$-dimensional psychological space and the distance between items $x^{(t)}$ and ξ^j is defined as follows:

$$d(x^{(t)}, \xi^j) = \left[\sum_{i=1}^{\mathfrak{N}} \omega_i^{(t)} \cdot |x_i^{(t)} - \xi_i^j|^r \right]^{\frac{1}{r}},$$

where $\omega_i^{(t)}$ is the attention weight associated to dimension i after the presentation of the t-th stimulus and r is a positive constant determined on the basis of the nature of the items (see Subsection 3.2.1 for further information about the choice of the constant r). Since both Experiment I and II involved highly separable-dimension stimuli, the value of r is set equal to 1.

Once the exemplar nodes are activated, these activations are multiplied by connection weights and summed to form the category outputs. More precisely, as a result of the reception of the t-th stimulus, the K category node ($K \in \{K_1, \ldots, K_N\}$) is activated by the following quantity:

$$O_K^{(t)} = \sum_{j=1}^{N_L} a_j^{(t)} \cdot w_{j,K}^{(t)}, \tag{3.26}$$

where $w_{j,K}^{(t)}$ is the weight of the connection that links exemplar node ξ^j to category node K at the t-th iteration (i.e., when the network receives the t-th stimulus).

Once the outputs for the categories are computed, the model computes the probability of classifying an item into a specif category as a function of the category outputs. Again, if the model is asked, it classifies stimuli on the basis of the computed probability. Similarly to Component-Cue, we considered both the exponential (Equation 3.18) and linear (Equation 3.19) versions of the classification probability. In the original paper [Kru92], the exponential version is described, while in subsequent studies both versions are employed [NKM92; Nos+94; Pal99]. Similarly to Component-Cue, the exponential version is denoted by an E (i.e., ALCOVE^E), while the linear version by an L (i.e., ALCOVE^L).

Once the classification probability of the input stimulus is computed, both the association and attention weights are updated. Similarly to Component-Cue, the updating of the weights is based on a gradient descent algorithm, whose aim is to minimize the difference between the outputs of the network and the feedback. The error of the network is defined as in Equation 3.21. Therefore, the updating of the association weights is the same as in Equation 3.23, while the updating of the attention weights is given by the following rule:

$$\begin{aligned}\omega_i^{(t+1)} &= \omega_i^{(t)} - \lambda_\omega \cdot \frac{\partial E^{(t)}}{\partial \omega_i^{(t)}} \\ &= \omega_i^{(t)} - \lambda_\omega \cdot \sum_{l=1}^{N} \sum_{j=1}^{N_L} a_j^{(t)} \cdot w_{j,K_l}^{(t)} \cdot c \cdot |x_i^{(t)} - \xi_i^j| \cdot \left(\mathcal{T}_{K_l}^{(t)} - O_{K_l}^{(t)}\right),\end{aligned} \tag{3.27}$$

where λ_ω is a positive freely estimated parameter representing the learning rate parameter for the attention weights. All weights (association and attention) are initiated at 0.

To recap, the parameters of the exponential version of ALCOVE are c, ϕ, λ_ω, λ_w, while those of the linear version are c, b, λ_ω, λ_w. Similarly to Component-Cue, ALCOVE can also be applied to reproduce transfer performance. To this purpose, the updating of both the association and attention weights is stopped and the static variant of ALCOVE is considered. The static variant of the model is denoted by ALCOVE-S^E (when the exponential version is considered) and ALCOVE-S^L (when the linear version is considered).

Example 3.9. Let us consider four learning exemplars (a black square, a gray square, a black circle and a gray circle) belonging to either category A or B. Figure 3.15 illustrates the structure of ALCOVE in this particular case. The intermediate nodes code the learning items (the black square, the gray square, the black circle and the gray circle), while the output nodes code the categories in with items are classified (A and B). ⊠

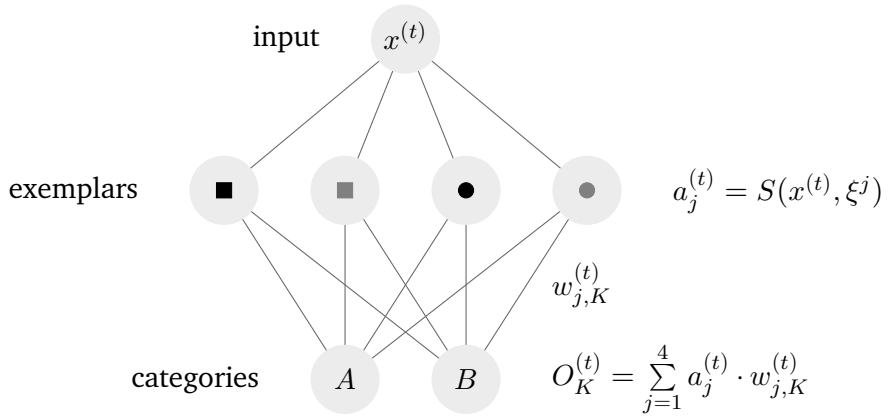


Figure 3.15 – *ALCOVE's neural network if the consider four exemplars: a black square, a gray square, a black circle and a gray circle.*

Remark 3.12. The Generalized Context Model (GCM) is included in the static variant of the linear version of ALCOVE (i.e., ALCOVE$^{\text{L}}$-S). Indeed, the GCM can be obtained from ALCOVE$^{\text{L}}$-S by setting:

$$w_{j,K} = \begin{cases} 1 & \text{if } \xi^j \in K \\ 0 & \text{otherwise} \end{cases}$$

for all $j \in \{1, \ldots, N_L\}$ and $K \in \{K_1, \ldots, K_N\}$. ⊠

Computational Aspects

The aim of this paragraph is to describe the way ALCOVE was coded. The following constraints were used:

i. The sensitive parameter c, the learning rate parameters of the association weights λ_w, and the learning rate parameters of the attention weights λ_ω were constrained to be greater than 0.

ii. The outputs was constrained to be between 1 and -1. Therefore, when the outputs were either greater than 1 or smaller than -1, the association weights $w_{j,K}^{(t)}$ ($j \in \{1, \ldots, \mathcal{N}\}$ and $K \in \{K_1, \ldots, K_N\}$) were re-scaled as follows:

$$w_{j,K}^{(t)} \leftarrow \frac{w_{j,K}^{(t)}}{|O_K^{(t)}|}.$$

This condition ensure that the classification probability of the linear version was positive (Equation 3.19). However, to maintain coherence between the two versions, the same constraint was also used in the exponential version.

iii. The attention weights were constrained to be positive and their sum was constrained to be equal to 1 (as in the GCM). Therefore, when the sum of the attention weights was not equal to 0, the attention weights $\omega_i^{(t)}$ $(i = 1, \ldots, \mathfrak{N})$ were re-scaled as follows:

$$\omega_i^{(t)} \leftarrow \max\left\{0, \frac{\omega_i^{(t)}}{\sum_{k=1}^{\mathfrak{N}} \omega_k^{(t)}}\right\}.$$

iv. The parameter b of the linear version was constrained to be grater than 1 to ensure that the classification probability in Equation 3.19 was positive.

TO SUM UP *Attention Learning COVEring Map Model (ALCOVE)*

ALCOVE is a learning model that integrates the exemplar-based representation of the GCM in a neural network structure. The structure of ALCOVE is composed of three layers: a single input node receiving the stimuli, a layer of exemplar nodes coding the learning items of the classification task, and a layer of category nodes coding the categories in which exemplars are classified. When a stimulus $x^{(t)}$ reaches the input node, the exemplar nodes ξ^j $(j = 1, \ldots, N_L)$ are activated according to the quantity:

$$a_j^{(t)} = S(x^{(t)}, \xi^j) = e^{-c \cdot \left[\sum_{i=1}^{\mathfrak{N}} \omega_i^{(t)} \cdot |x_i^{(t)} - \xi_i^j|^r\right]^{\frac{p}{r}}},$$

where $S(x^{(t)}, \xi^j)$ denotes the similarity between exemplars $x^{(t)}$ and ξ^j, c is a sensitive parameter, p and r are positive constants, and $\omega_i^{(t)}$ is the attention weight relating to dimension i at the t-th iteration. The activations of the exemplar nodes are multiplied by association weights and summed to form outputs. The K category node is activated by the quantity:

$$O_K^{(t)} = \sum_{j=1}^{N_L} a_j^{(t)} \cdot w_{j,K}^{(t)},$$

where $w_{j,K}^{(t)}$ is the weight of the connection that links exemplar node ξ^j to category node K at the t-th iteration ($K \in \{K_1, \ldots, K_N\}$).

Finally, the probability of classifying stimulus $x^{(t)}$ into category $K \in \{K_1, \ldots, K_N\}$ is given by:

$$\mathbb{P}\left(K \mid x^{(t)}, \mathcal{H}_{t-1}\right) = \frac{e^{\phi O_K^{(t)}}}{\sum_{j=1}^{N} e^{\phi O_{K_j}^{(t)}}} \qquad \text{or} \qquad \frac{O_K^{(t)} + b}{\sum_{j=1}^{N} \left(O_{K_j}^{(t)} + b\right)}$$

depending whether the exponential or linear version is considered. The quantities ϕ and b are two positive freely estimated parameters.

Both the association and attention weights are updated using a gradient descent algorithm to minimize the difference between the outputs of the model and the feedback. The weights are updated using the following learning rules ($j \in \{1, \ldots, N_L\}$ and $K \in \{K_1, \ldots, K_N\}$):

$$w_{j,K}^{(t+1)} = w_{j,K}^{(t)} + \lambda_w \cdot a_j^{(t)} \cdot \left(\mathcal{T}_K^{(t)} - O_K^{(t)}\right),$$

$$\omega_i^{(t+1)} = \omega_i^{(t)} - \lambda_\omega \cdot \sum_{l=1}^{N} \sum_{j=1}^{N_L} a_j^{(t)} \cdot w_{j,K_l}^{(t)} \cdot c \cdot |x_i^{(t)} - \xi_i^j| \cdot \left(\mathcal{T}_{K_l}^{(t)} - O_{K_l}^{(t)}\right),$$

where λ_w and λ_ω are positive freely estimated parameters and

$$\mathcal{T}_K^{(t)} = \begin{cases} 1 & \text{if } x^{(t)} \in K \\ -1 & \text{otherwise.} \end{cases}$$

Both the association and attention weights are initiated at 0. To recap, the parameters of the model are

$$c,\ \phi,\ \lambda_\omega,\ \lambda_w \qquad \text{or} \qquad c,\ b,\ \lambda_\omega,\ \lambda_w$$

depending whether the exponential or linear version is considered. ALCOVE can also be applied to reproduce transfer performance stopping the updating of both the association and attention weights. We refer to this variant of the model as the static variant.

3.3.3 Relations Between Learning Models

Although Component-Cue and ALCOVE share a similar structure based on neural networks, they integrate two distinct strategies: a rule-based strategy for Component-Cue and a similarity-based strategy for ALCOVE. Component-Cue aims to determine the set of features that are a good predictor of the category membership of the items. In other words, it induces the simplest rule on the basis of the features of the stimuli. For example, if the classification task consists in classifying a series of pictures (a tuna, a catfish, an eagle, a pigeon, a penguin etc.) into either the category of fishes or the category of birds, Component-Cue would predict that the ovoid shape is a good predictor of the category of fishes while the ovoid shape with wigs is a good predictor of the category of birds. However, using this strategy it would probably fail when pictures of penguins are considered.

Conversely, ALCOVE uses a similarity strategy to determine the category membership of the items. When a stimulus is presented, ALCOVE classifies it on the basis of its similarity to the learning items. Moreover, it is also able to memorize individual examples by tuning the sensitive parameter. Therefore, in the previous example, ALCOVE would classify the pictures of tuna and catfish into the category fishes, and the pictures of eagle and pigeon into the category birds because of their similarities. However, by increasing the sensitive parameter, ALCOVE would learn to classify items by rote (even close items would be perceived as dissimilar). Thus, when pictures of penguins are presented, ALCOVE would have learned by heart that penguins are members of the category birds.

One of the main advantage of Component-Cue is its ability to induce the simplest rule. However, the binary-valued activation of its nodes (a node is activated with a value of either 0 or 1) and its inability to learn some category membership by rote limit its application to a small number of categories (e.g. categories based on a “principal rule” structure). Conversely, ALCOVE is adapted to a larger variety of categories because of the continuous-valued activation of its nodes (a node is activated as a function of its similarity to the input stimulus) as well as the integration of an attention mechanism.

4

Advanced Inference Method and Application to Transfer Models

Contents

There is a variety of practices that has been used in the literature to fit categorization models to experimental data. A number of studies have relied on the use of computer simulations [CG19; NSM17]. The overall predictions of the model were obtained by averaging the classification predictions generated through simulations. Other studies have used the same set of observations to both estimate the parameters of the model and compute the predictions [NSM17; Nos+18; SN20]. Finally, different research have used different techniques to estimate the parameters of the model (e.g., SSD, likelihood trial-by-trial, likelihood block-by-block, likelihood epoch-by-epoch, etc.) [NKM92; Nos+94;

Nos+18]. We believe that the field of categorization would benefit from the use of a single and robust inference method.

Goals

The first goal of this chapter is to provide a general and robust inference method to both fit categorization models to experimental data and compare them. Although this method is not specific to cognitive models and uses generic statistical tools, the reasons that brought us to adopt it have been dictated by the characteristics of the studied models.

The second goal of this chapter is to use the previous inference method to compare the selected transfer models and determine the model that best fits the data. The transfer models were compared on the transfer phase of Experiment I.

Outline of this chapter

Firstly, we describe some fundamental statistical concepts such us underfitting, overfitting, and bias-variance trade-off. Secondly, we provide a visual representation of the predictions of the transfer models. Thirdly, we describe how the parameters of the models were estimated and evaluate the consistency of the estimates as a function of the size of the dataset. Then, we detail the cross-validation techniques that were used to fit models to data (one adapted to transfer models, the other to learning models). Finally, we apply one of the cross-validation technique (the one adapted to transfer models) to determine the transfer model that best describes the transfer phase of Experiment I.

4.1 Preliminaries

This section represents an introduction to some fundamental statistical concepts such as underfitting, overfitting, and bias-variance trade-off. It has been specifically conceived for readers that are not familiar with these concepts. Let us start with an example. Let us suppose that one measured the weight and height of a group of elephants (see Figure

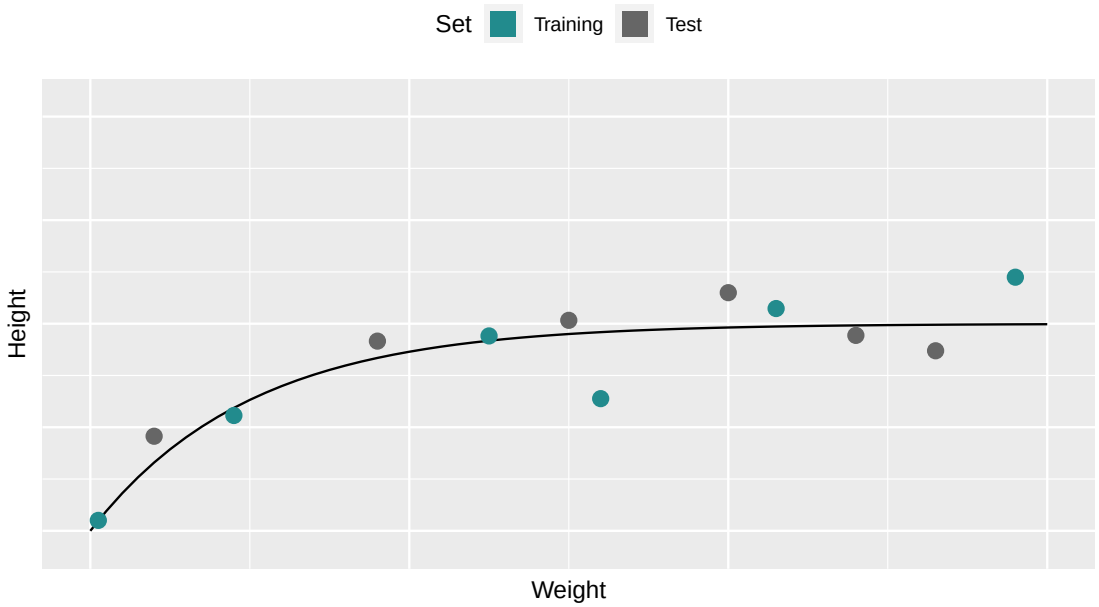


Figure 4.1 – *Plot of the weight and height of a group of elephants. The line represents the function that describes (with the addition of a noise) the relation between weight and height. The blue dots represent the training set on which the models are trained, while the gray dots represent the testing set on which the models are evaluated.*

4.1). Ideally, there is a function f that describes the relation between weight and height (the line in the graph), with the addition of a noise:

$$\text{height} = f(\text{weight}) + \text{noise}.$$

However, the function f is unknown. A common way to approximate this relation is to use a model. Here, three models are considered: three polynomials with, respectively, one, three, and six free parameters.

A statistical technique that allows one to fit models to data as well as evaluate their performance consists in splitting the data into two sets, one for training the models and the other for testing them (i.e., cross-validation). In Figure 4.1, the blue and gray dots represent the training and testing sets, respectively. For each model, the training set is used to find the parameters that best fit the training points. For instance, training the polynomial with one free parameter consists in identifying the horizontal line that minimizes a specific criterion (in our case, the sum of squared deviations) on the training points. For each model, the best fit on the training points represents the best approximation of the true relation achieved by the model.

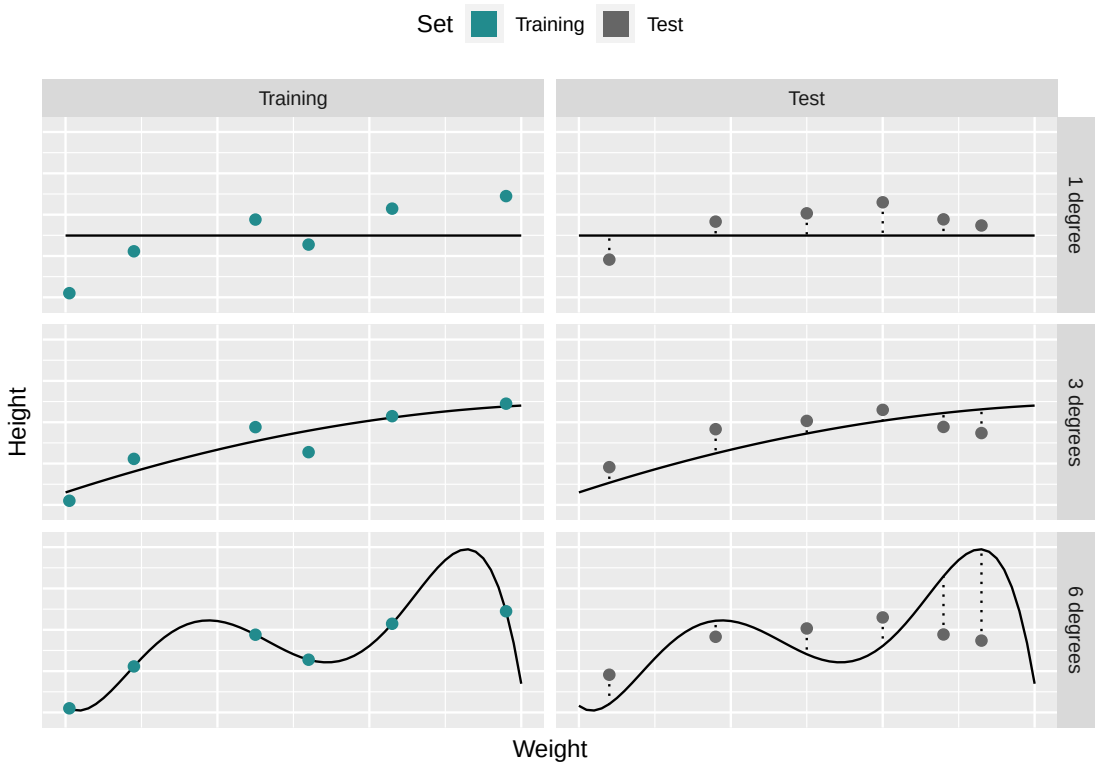


Figure 4.2 – *Training and testing of three polynomial models having one, three and six free parameters, respectively (in the graph the free parameters are referred as degrees of freedom). The blue dots are used to train the models while the gray dots are used to evaluate them. Testing errors on the training and testing sets are shown in Table 4.1.*

Once the approximation of the true function is found, one needs to evaluate the performance of the model on the testing set using a specific criterion (in our case, the sum of squared deviations). The evaluation of the models performed using the sum of squared deviations is usually called testing error. For instance, in the case of the polynomial with one free parameter, its testing error is equal to the sum of the squared distances between the test points and the trained horizontal line.

Naively, we could expect that the model with the highest number of parameters performs better on both the training and testing sets. Let us verify this expectation. The results of the training and testing of the three polynomial models are illustrated in Figure 4.2, while their training and testing errors are provided in Table 4.1. As we expected, the higher the complexity of the model (the number of free parameters), the smaller the training error. Intuitively, the higher the number of free parameters, the better a model

	1 degree	3 degrees	6 degrees
Training error	0.9	0.1	0
Testing error	0.4	0.3	2

Table 4.1 – *Training and testing errors that the three polynomials with one, three and six free parameters made on the training and testing sets, respectively.*

can adapt itself to fit a set of points. For example, in the case of the polynomial with six free parameters, it is always possible to find a curve passing through six training points, therefore its training error is equal to zero.

Let us now look at the testing error. Is it what we expected? Partially yes. In fact the polynomial with three parameters did better than the one with one parameter. Since the data follow a curvy line, the horizontal line is not complex enough to replicate the curvy behavior that underlies the true relation. This phenomenon is called underfitting. However, if we look at the performance of the polynomial with six parameters, the result does not meet our expectation. Indeed, the complex model performed worse than the model with one parameter. This raises the questions: why did the complex model perform so badly on the testing set? Why did it perform even worse than the simplest model? The answer lies in the noise. Since the data are characterized by a certain amount of noise, the complex model detected in the noise patterns that do not exist, which caused its failure on the testing set. This phenomenon is called overfitting. Let us properly summarize the phenomena of underfitting and overfitting.

Underfitting. This phenomenon occurs when the complexity of the model is too low as compared to the complexity of the to-be-estimated function. This discrepancy in complexity produces the inability for the model to entirely capture the underlying pattern of the data. Models with fewer parameters tend to underfit the data.

Overfitting. This phenomenon occurs when a model captures the noise of the data along with their underlying structure. Models with higher parameters tend to overfit the data.

The testing errors in Table 4.1 can be decomposed into two separate sources of error. Indeed, the expected squared error between the true relation and the approximated one can be decomposed into bias and variance [GBD92], [KW97], [LS08], [Mun+10] (we

refer the readers to Box 4.1 for a mathematical description of the bias-variance error decomposition).

Bias. The bias represents the difference between the average prediction of the model and the value determined by the true relation. This type of error is caused by too simplistic model assumptions as compared to the complexity of the true function. Simple models that are unable to capture the underlying patterns of the data entirely (underfitting) tend to have a high bias. Thus, on one hand, models with a high bias do not produce great predictions but only good ones. However, on the other hand, their predictions are consistently good. In other words, the quality of their predictions is moderately affected when a new set of training points is considered.

Variance. The variance indicates the variability of the predictions of the model. This type of error is due to either the large amount of noise in the data, or to the limited size of the training set. The higher the amount of noise in the data, the higher the variance of the model. Inversely, the smaller the size of the training data, the higher the variance of the model. Models with a higher complexity as compared to that of the true function tend to have a high variance. Models with high variance are more likely to capture the noise in the training data (overfitting), which leads to great performance on training data but highly variable performance on test data (predictions could be good sometimes and bad other times).

Ideally, a good model would have low bias, capturing all the relevant information in the data, and low variance, avoiding to detect patterns in the noise. However in general, models with low bias have high variance whereas models with low variance have high bias. This impossibility to minimize both errors simultaneously is called the bias-variance trade-off. Therefore, a good model is characterized by balanced bias-variance errors. The polynomial with three parameters presented in the example above represents a balanced model.

There are two main approaches to select the model that both provides a good account of the data and is characterized by balanced bias-variance errors. The first approach consists in estimating the parameters of the model and testing them on different sets of observations. Examples include the hold-out, the k-fold, and the leave-p-out methods (i.e., cross-validation methods). An alternative approach to model selection involves using probabilistic statistical measures that attempt to quantify both the complexity of the model and its performance on the training dataset. Examples include the Akaike

and Bayesian Information Criterion (AIC and BIC), and the Minimum Description Length (MDL). To conclude, a commonly accepted principle is the Occam's Razor (also called principle of parsimony) which states that "one should not increase, beyond what is necessary, the number of entities required to explain anything" (see also [PM02]).

Considerations about the studied categorization models

Let us clarify two important points that could increase the risk that the selected categorization models overfit the data. The first point is that categorization models (learning models in particular) have such a complex structure that their number of parameters does not necessarily correspond to the dimension of the space of their attainable predictions. The second point is that model predictions are expressed in terms of probability (the probability of classifying stimuli into category A or B), whereas the data are expressed in terms of binary responses (1 when participants classified stimuli into category A, 0 otherwise). Therefore, the inherent noise in the data can be very high, especially when the model predicts the classification probability to be in the surroundings of 0.5. To recap, the selected categorization models are at risk of overfitting the data. In the next section, we describe the approach we adopted to limit the risk of overfitting the data.

Box 4.1 *Bias-Variance Error Decomposition*

Let us assume that an independent variable X affects the value of a dependent one Y via the following formula:

$$Y = f(X) + \epsilon,$$

where f is an unknown function and ϵ is a random variable. Let us suppose that we have a training set D_n consisting of a set of points $x_1, \ldots, x_n$ as well as their associated values $y_1, \ldots, y_n$:

$$D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}.$$

By means of the training set D_n and some learning algorithm, we find a function $\hat{f}_{D_n}(x)$ that approximates as well as possible the function $f(x)$ in the training

points D_n. The expected squared difference between the real function and the approximated one on an unseen sample x can be decomposed as follows:

$$\begin{aligned}
&\mathbb{E}_{D_n}\left[\left(f(x)-\hat{f}_{D_n}(x)\right)^2\right] \\
&\qquad = \left(\mathbb{E}_{D_n}\left[\hat{f}_{D_n}(x)\right]-f(x)\right)^2+\mathbb{E}_{D_n}\left[\hat{f}_{D_n}(x)^2\right]-\mathbb{E}_{D_n}\left[\hat{f}_{D_n}(x)\right]^2 \\
&\qquad = \text{Bias}(\hat{f})^2+\text{Var}(\hat{f}).
\end{aligned}$$

Let us detail the proof of the above identity. For convenience, we abbreviate $f(x)$ with f, $\hat{f}_{D_n}(x)$ with $\hat{f}$ and we drop the D_n subscript on our expectation operators. We have that:

$$\begin{aligned}
\mathbb{E}\left[\left(f-\hat{f}\right)^2\right] &= \mathbb{E}\left[\left(f-\hat{f}+\mathbb{E}[\hat{f}]-\mathbb{E}[\hat{f}]\right)^2\right] \\
&= \mathbb{E}\left[\left(f-\mathbb{E}[\hat{f}]\right)^2\right]+\mathbb{E}\left[\left(\mathbb{E}[\hat{f}]-\hat{f}\right)^2\right] \\
&\qquad +2\,\mathbb{E}\left[\left(\mathbb{E}[\hat{f}]-\hat{f}\right)\left(f-\mathbb{E}[\hat{f}]\right)\right] \\
&= \left(f-\mathbb{E}[\hat{f}]\right)^2+\mathbb{E}\left[\left(\mathbb{E}[\hat{f}]-\hat{f}\right)^2\right] \\
&\qquad +2\,\mathbb{E}\left[\mathbb{E}[\hat{f}]-\hat{f}\right]\left(f-\mathbb{E}[\hat{f}]\right) \\
&= \left(f-\mathbb{E}[\hat{f}]\right)^2+\mathbb{E}\left[\left(\mathbb{E}[\hat{f}]-\hat{f}\right)^2\right] \\
&= \text{Bias}(\hat{f})^2+\text{Var}(\hat{f}).
\end{aligned}$$

In the previous proof we used the following two facts:

i. Since $f-\mathbb{E}[\hat{f}]$ is deterministic, then $\mathbb{E}\left[\left(f-\mathbb{E}[\hat{f}]\right)^2\right]=\left(f-\mathbb{E}[\hat{f}]\right)^2$.

ii. $\mathbb{E}\left[\mathbb{E}[\hat{f}]-\hat{f}\right]=\mathbb{E}[\hat{f}]-\mathbb{E}[\hat{f}]=0$.

4.2 Visual Representation of Models

The selected models are very complex, either because of their intricate structure, or because of their high number of parameters, or even both. These characteristics lead to two main issues.

On one hand, the intricate structure makes unclear whether or not the number of parameters reflects the dimension of the space of their attainable predictions. For instance, in the OGCM, the integration of the ordinal dimension might allow distinct parameters to generate identical predictions. Without this information it is an issue to determine whether the spaces of their predictions overlap or not. Incidentally, this issue represents the reason why it is not advisable to use criteria as the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), or equivalent criteria to select the model that best describes the data. Since these criteria penalize models on the basis of their number of parameters, there is a high risk to unfairly penalize some models (this concept will be better described in Section 4.4).

On the other hand, the large number of parameters can result in identifiability issues. An example is represented by both the linear and exponential versions of ALCOVE-S. The fact that the number of parameters of these models exceeds the dimension of the search space (i.e. the number of learning and transfer items) makes the function associating the parameters with the probabilities non-injective. Identifiability issues can have a great impact on the estimation of the parameters of the models (as we will see in Section 4.3). In view of the above, we considered that a preliminary visualization of the models was warranted.

Models generate different classification probabilities, depending on the values of their free parameters. Yet, there could be probability patterns that are attainable for some models but not for others. For instance, patterns in which the category A items are classified into category A with a low probability can only be reached by the Generalized Context Model in highly limited border cases. Indeed, the probability of classifying a category A item into category A takes the form $\frac{1+\epsilon_A}{1+\epsilon_A+\epsilon_B}$, where ϵ_A represents the summed similarities between the considered stimuli and the other category A stimuli and ϵ_B represents the summed similarities between the considered stimuli and the other category B stimuli. The two quantities ϵ_A and ϵ_B have usually the same magnitude, leading the classification probability to be greater than 0.5.

The aim of this section is to investigate the way the spaces of the predictions of the models are interconnected. Although the static variants of the learning models were not used to analyze the transfer phase of Experiment I, they have been nonetheless included in this analysis.

4.2.1 Principal Component Analysis (PCA)

The description of the procedure is given for transfer models (learning models will be studied in Chapter 5). Let $M \in \mathfrak{M}$ be a transfer model and $\theta_M \in \Theta_M$ its set of parameters. The items of the classification task (training and transfer) are denoted by $\xi_1, \ldots, \xi_N$. Given a set of parameters, the probability that the model classifies stimuli $\xi_1, \ldots, \xi_N$ into category A can be computed. Thus, each set of parameters θ_M can be associated with a vector including the probability that the model classifies stimuli $\xi_1, \ldots, \xi_N$ into category A as follows:

$$\begin{aligned} g_M : \Theta_M &\longrightarrow [0,1]^N \\ \theta_M &\longmapsto P_M^{\theta_M} = \left(\mathbb{P}_M^{\theta_M}(A \,|\, \xi_1), \ldots, \mathbb{P}_M^{\theta_M}(A \,|\, \xi_N)\right). \end{aligned} \tag{4.1}$$

The vector $P_M^{\theta_M}$ is called the probability pattern associated with the model M and the value θ_M (we underline that the probability patterns also depend on the categories of the classification tasks). The spaces of the predictions of the models can be studied by analyzing the image $g_M(\Theta_M)$ of each model $M \in \mathfrak{M}$. Let us describe the steps of this analysis.

Step #1

The first step consists in randomly choosing l sets of parameters for each model $M \in \mathfrak{M}$. The selected sets of parameters are denoted by $\theta_M^1, \ldots \theta_M^l$ (for each $M \in \mathfrak{M}$).

Step #2

The second step consists in computing (for each model M) the probability patterns $P_M^{\theta_M^i} = g_M(\theta_M^i)$ associated to the selected set of parameters θ_M^i ($i = 1, \ldots, l$).

Step #3

The third and last step consists in applying the Principal Component Analysis technique (see Box 2.9) to the table composed of the $l \times |\mathfrak{M}|$ probability patterns. Each probability

pattern $P_M^{\theta_M^i}$ (for $i = 1, \ldots, l$ and $M \in \mathfrak{M}$) represents a row of the table. This step allows the visualization of the predictions of the models on a 2D-plan.

PROCEDURE SUMMARY 4.1 *Principal Component Analysis (PCA)*

Objective: To investigate the probability patterns (i.e., the probability of classifying stimuli into category A) that are attainable for the categorization models.

#1. Randomly select a choice of sets of parameters for each of the considered models.

#2. For each considered model and for each set of parameters, compute the probability pattern associated to the model and the set of parameters (i.e. the probability of classifying the items into category A, given the model and the specific set of parameters).

#3. Apply the Principal Component Analysis (PCA) to the table composed of all the probability patterns.

4.2.2 Simulated Transfer Data Analysis

In this subsection, the procedure previously described (Procedure Summary 4.1) is applied to both the transfer models and the static versions of the learning models.

Technical Aspects

i. The analysis included the following models: the GCM, the GCM-Lag, the three versions of the OGCM (OGCM-T, OGCM-L, and OGCM-M), both the linear and exponential versions of the static variant of the Component-Cue model (Component-Cue$^{\text{L}}$-S and Component-Cue$^{\text{E}}$-S), and both the linear and exponential versions of the static variant of ALCOVE (ALCOVE$^{\text{L}}$-S and ALCOVE$^{\text{E}}$-S).

ii. The studied categories were the 5-4 category set of Experiment I (see Figure 2.1). Moreover, the probability pattern included the classification probability of all training and transfer items of Experiment I ($N = 16$).

iii. A total of 20 000 sets of parameters per model was considered (i.e., $l = 20\,000$). Each set was randomly chosen among those respecting the constraints of the models.

Results

Figure 4.3a and 4.3b show three different planes of the PCA resulting from the application of Procedure Summary 4.1 to the considered models. The linear versions of the learning models produced similar results to those obtained considering the exponential versions. Therefore, the linear versions were not included in the graphs to avoid their overburdening. The position of a set of relevant probability patterns is highlighted in both figures to facilitate the interpretation of the graphs. The highlighted probability patterns are the following (a visual description of the patterns is equally given in Figure 4.3a, on the bottom; the word "pattern" is omitted to avoid cluttering):

Pattern A. Both learning and transfer items are classified into category A, regardless of their effective category.

Pattern B. Both learning and transfer items are classified into category B.

Pattern C. Learning items are classified into their effective category (i.e., category A items into category A and category B items into category B), while transfer items are associated with a probability of 0.5 (i.e., random classification).

Pattern D. Learning items are classified into their opposite category (i.e., category A items into category B and category B items into category A), while transfer items are associated with a probability of 0.5 (i.e., random classification).

Pattern E. Both learning and transfer items are randomly classified. They are both associated with a probability of 0.5.

Pattern F. Learning items are classified into their effective category (i.e., category A items into category A and category B items into category B), while transfer items are classified into category A.

Pattern G. Learning items are classified into their effective category (i.e., category A items into category A and category B items into category B), while transfer items are classified into category B.

Pattern H. Learning items are classified into their effective category (i.e., category A items into category A and category B items into category B), while transfer items are classified according to the rule-based classification (i.e., all gray items into category A, while all blue items into category B).

Pattern I. Learning items are classified into their effective category (i.e., category A items into category A and category B items into category B), while transfer items are classified according to the similarity-based classification (i.e., the classification is established on the basis of the category membership of the closest items).

The analysis of the planes of the PCA showed four groups of models manifesting a similar behavior.

First group. The first group of models included the static variants of the learning models (i.e., the Component-Cue^{E}-S, the Component-Cue^{L}-S, the ALCOVE^{E}-S, and the ALCOVE^{L}-S). These models seemed to be capable to attain every probability pattern. Indeed, they generated a variety of patterns ranging from patterns in which all items are classified into a single category (i.e., patterns A and B) to patterns in which learning items are either classified into their effective category (i.e., patterns C, F, G and H), or into their opposite category (i.e., pattern D). The large number of parameters of these models produced highly variable probability patterns.

Second group. The second group only included the GCM-Lag. This model occupied a smaller area as compared to the previous models. The attainable patterns of the model seemed to be those in which learning items are correctly classified (i.e., the correct probability is greater than 0.5). Indeed, the GCM-Lag is unable to reach patterns in which learning items are incorrectly classified (i.e., the triangle between patterns A, B and D). Additionally, the patterns generated by the GCM-Lag always included those generated by the GCM (which is plausible since the GCM-Lag is an extension of the GCM).

Third group. The third group included both the GCM and the OGCM-T. These models were capable to attain a smaller variety of patterns as compared to the GCM-Lag. The spreading of their patterns was restrained to patterns in which learning items are correctly classified (i.e., the correct probability is greater than 0.5) and transfer items are randomly classified. The GCM and the OGCM-T were unable to reach either patterns in which learning items are incorrectly classified (i.e., pattern D), or patterns in which learning items are correctly classified and transfer items are classified in one single category (i.e., patterns F and G). Additionally, the patterns

generated by the OGCM-T always included those generated by the GCM (which is plausible since the OGCM-T is an extension of the GCM).

Forth group. The forth group included both the OGCM-L and the OGCM-M. Their patterns were similar to those of the previous group (see Figure 4.3a and 4.3b, on the top). However, in some planes of the PCA these models reached a larger variety of patterns (see Figure 4.3b, on the bottom). Additionally, the patterns generated by both the OGCM-L and OGCM-M always included those generated by the GCM (which is plausible since they are both an extension of the GCM).

The PCA analysis led us to the conclusion that the connection between the number of parameters of the studied models and the dimension of the space of their predictions is not straightforward. For instance, the GCM-Lag and the OGCM-T have the same number of parameters but the spreading of their patterns is different.

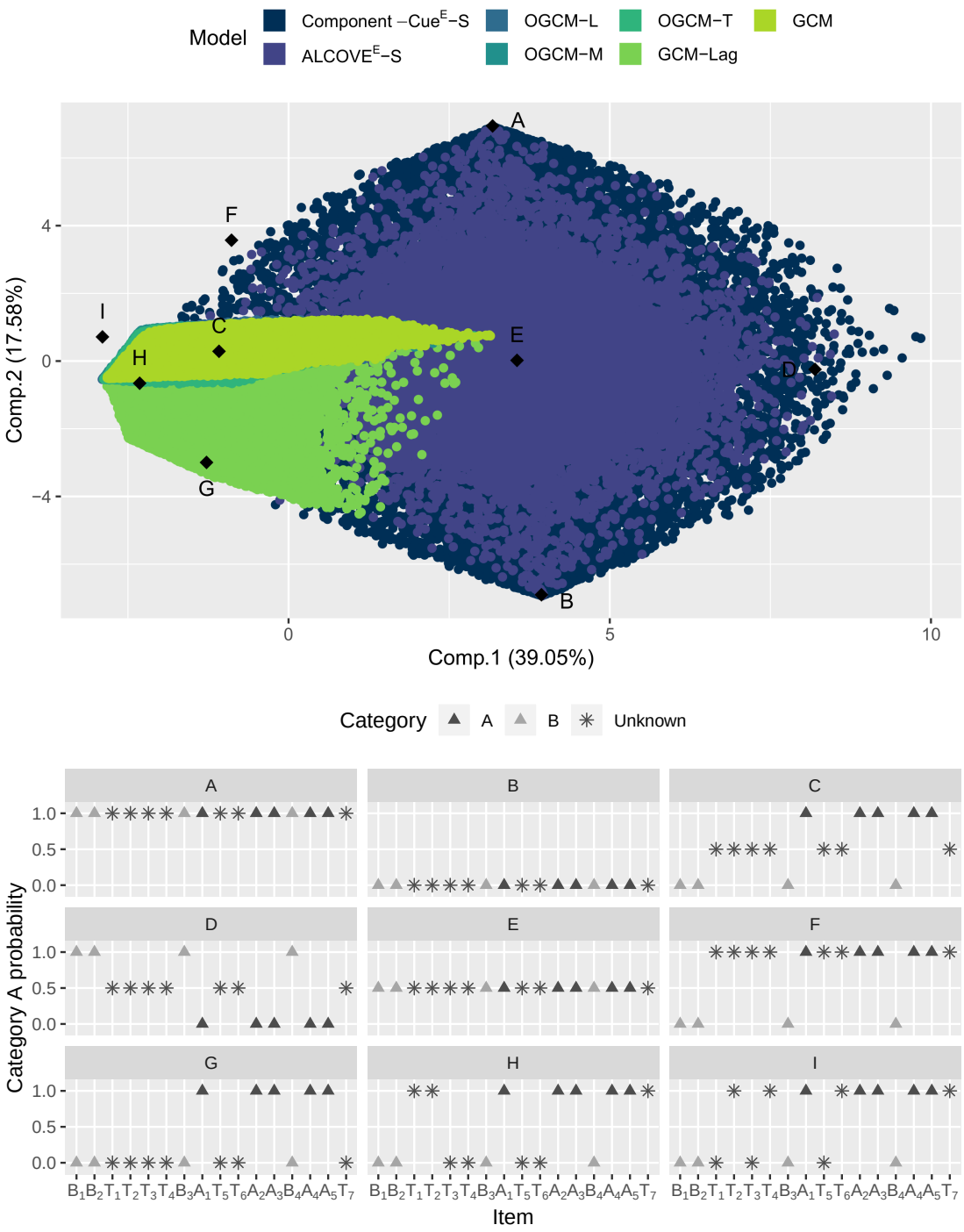


Figure 4.3a – *Result of the Principal Component Analysis (PCA) applied to probability patterns of the studied transfer models. The patterns refer to the 5-4 category set of Experiment I (transfer items included). On the top, the projection of the probability patterns on the first and second components (there is a total of 20 000 patterns per model). On the bottom, some relevant probability patterns.*

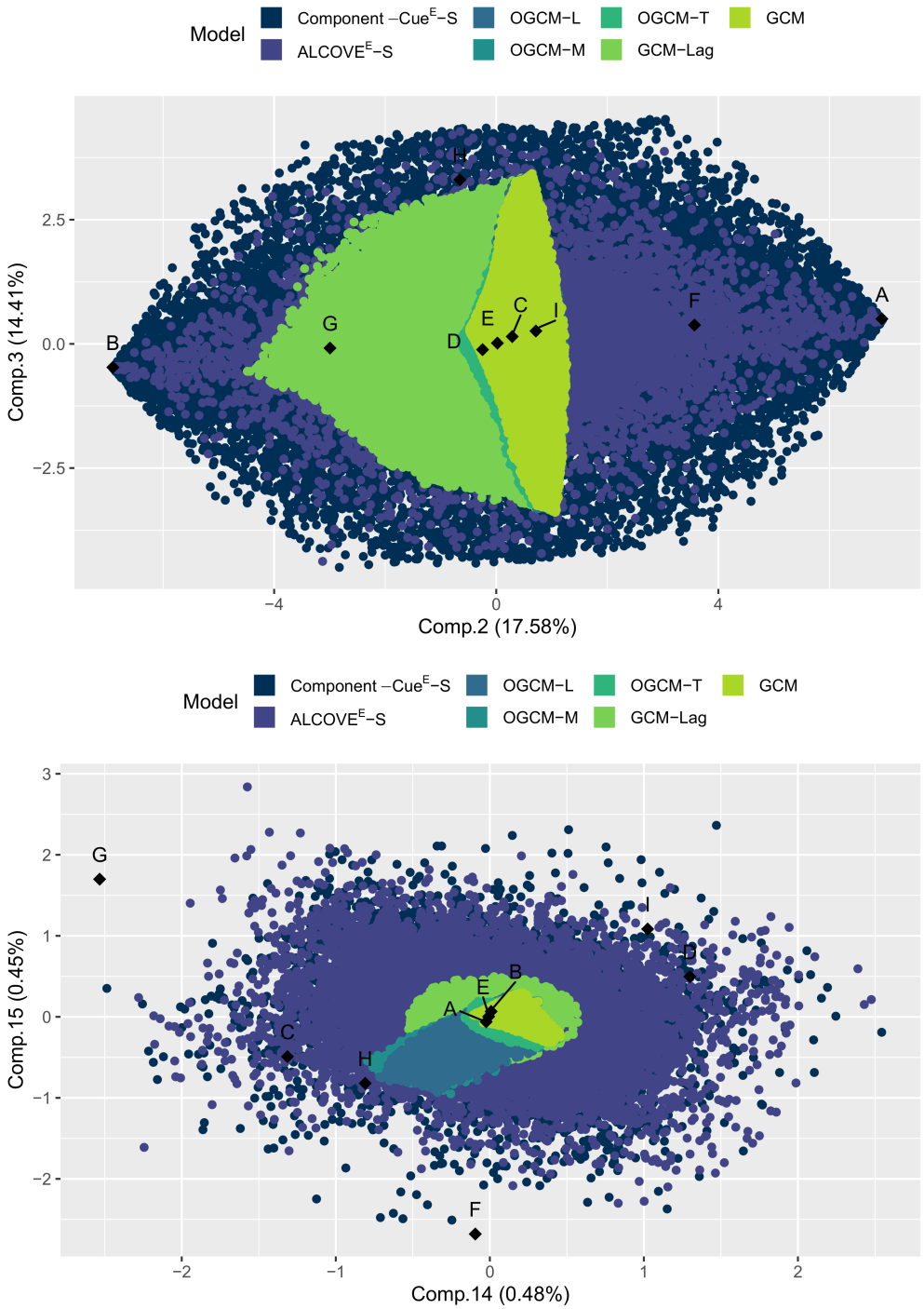


Figure 4.3b – *Result of the Principal Component Analysis (PCA) applied to probability patterns of the studied transfer models. The patterns refer to the 5-4 category set of Experiment I (transfer items included). On the top, the projection of the probability patterns on the second and third components. On the bottom, the projection on the fourteenth and fifteenth components. There is a total of 20 000 patterns per model.*

4.3 Parameter Estimation

The Maximum Likelihood Estimation (MLE) is the method we adopted to estimate the parameters of the models. Under certain conditions (including the identifiability) the maximum likelihood estimator is consistent. However, not all the studied models are identifiable (as seen in the introduction to the previous section). For instance, the OGCM could generate the same probability patterns for two distinct sets of parameters because of the interference of the presentation order. This issue requires us to evaluate the consistency of the estimates using computer simulations. Firstly, we briefly describe the MLE procedure, and secondly, we examine how to validate the consistency of the MLE on simulated data.

4.3.1 Maximum Likelihood Estimation (MLE) by Gradient Descent

The method that was used to estimate the parameters of the models is the Maximum Likelihood Estimation (MLE) [Ald97]. This technique consists in estimating the parameters of the models by maximizing its likelihood, evaluated on some observed data. Again, the likelihood measures the goodness-of-fit of a model to the observed data as a function of its parameters. Therefore, the parameters that maximize the likelihood are the parameters with which the model is more likely to generate the observed data.

Remark 4.1. To find the parameters that maximize the likelihood is equivalent to find the parameters that minimize the opposite of the logarithm of the likelihood (the logarithm is a monotonically increasing function).

$$\hat{\theta} \in \arg\max_{\theta \in \Theta} \mathcal{L}_M\left(D\,;\,\theta\right) \iff \hat{\theta} \in \arg\min_{\theta \in \Theta} \left\{-\log \mathcal{L}_M\left(D\,;\,\theta\right)\right\}$$

However, from a computational point of view, minimizing the opposite of the logarithm of the likelihood is easier than maximizing the likelihood. Indeed, *i)* there are a larger number of algorithms implementing function minimization rather than function maximization, and *ii)* the application of the logarithm transforms the products in the likelihood (see Subsection 3.1.1) into sums, which are more convenient. Therefore, the minimizing the opposite of the logarithm of the likelihood was preferred to the maximization of the likelihood. ⊠

To find the parameters that maximize the likelihood of the models we used the procedure described below.

PROCEDURE SUMMARY 4.2 *MLE by means of the Gradient Descent Algorithm*

Objective: To find the set (or sets) of parameters $\hat{\theta} \in \Theta$ that minimizes the opposite of the logarithm of the likelihood of a model M, evaluated on some observed data D:

$$\hat{\theta} \in \underset{\theta \in \Theta}{\arg\min} \left\{ -\log \mathcal{L}_M(D\,;\,\theta) \right\}.$$

#1 Minimize the opposite of the logarithm of the likelihood (evaluated on some observed data) using the gradient descent algorithm (see Box 4.2). The gradient descent algorithm is initialized at a random initial condition.

#2 Recover the parameters associated to the minimum.

#3 Iterate Step #1 and #2 several times considering different initial conditions for the gradient descent algorithm to avoid local minima. Let us denote by $\hat{\theta}_1, \ldots, \hat{\theta}_l$ the parameters associated with the maxima resulting from the application of the gradient descent algorithm l times.

#4 Select the smallest minimum and determine the set of parameters $\hat{\theta}$ associated with the latter:

$$\hat{\theta} \in \underset{i=1,\ldots,l}{\arg\min} \left\{ -\log \mathcal{L}_M(D\,;\,\hat{\theta}_i) \right\}.$$

BOX 4.2 *Gradient Descent Algorithm*

Gradient descent is a first-order iterative optimization algorithm for finding the local minimum of a differentiable function. The way this algorithm approaches a local minimum is by taking steps proportional to the negative of the gradient of the function at the current point. Since the gradient gives the direction of the highest increase of the function, following the negative of the gradient allows approaching a local minimum. Gradient descent was originally proposed by Cauchy in 1847.

Let us briefly illustrate how the gradient descent algorithm works. Let us consider a differentiable function $f(x)$. The algorithm starts with a guess x_0 for the local minimum of f and builds up the sequence $x_0, x_1, x_2, \ldots$ such that

$$x_{i+1} = x_i - \lambda_i \nabla f(x_i),$$

for all $i \geq 0$. The convergence of the sequence to a local minimum is strictly dependent on the value of the constants $\lambda_i \in \mathbb{R}_+$. With certain assumptions on the function f (for example convex or Lipschitz) and particular choices of λ_i, the convergence is guaranteed.

4.3.2 Validation of the Maximum Likelihood Estimation

As seen in the introduction to this section, the consistency of the MLE is not guaranteed when models are not identifiable. In this subsection, we describe the procedure used to express the accuracy of the parameter estimation as a function of the size of the dataset (the description is given for transfer models). The procedure is composed of the following steps.

Step #1

Let $M \in \mathfrak{M}$ be a transfer model. The first step consists in randomly choosing a value $\theta \in \Theta$ for the parameters of the model M (this value will be used to generate the observations on which the estimation will be performed).

Step #2

The second step consists in generating the observations used for the estimation, given the model and the choice of parameters. Let us assume that the simulated dataset is of size n. Thus, a sequence of n stimuli $x^{(i)} \in \{\xi_1, \ldots, \xi_N\}$ is considered. Each stimulus is selected from a set of learning and transfer items denoted by $E = \{\xi_1, \ldots, \xi_N\}$. Using

the sequence of stimuli, n observations $z^{(1)}, \ldots, z^{(n)}$ are generated, given the model M and the value θ. Each observation $z^{(i)}$ is a realization of $Z^{(i)}$ such that

$$Z^{(i)} \sim \mathcal{B}\left(\mathbb{P}_M^{\theta}\left(A \,|\, x^{(i)}\right)\right).$$

The observations $z^{(1)}, \ldots, z^{(n)}$ represent the simulated sequence of responses.

Step #3

The third step consists in performing the maximum likelihood estimation (Procedure Summary 4.2) on the previously generated observations to find the parameters $\hat{\theta}$ that satisfy the following condition:

$$\hat{\theta} \in \underset{\theta \in \Theta}{\arg\min} \left\{ -\log \mathcal{L}_M(z^{(1)}, \ldots, z^{(n)}\,;\, \theta) \right\}.$$

Step #4

The last step consists in computing the relative error between the generator parameter and the estimated parameter,

$$\frac{\theta - \hat{\theta}}{|\theta|},$$

as well as the error between the probability pattern computed with the generator parameter and the one computed with the estimated parameter,

$$\mathbb{P}_M^{\theta}\left(A \,|\, \xi_i\right) - \mathbb{P}_M^{\hat{\theta}}\left(A \,|\, \xi_i\right),$$

for $i = 1; \ldots, N$. This procedure is iterated with different values of n and θ.

PROCEDURE SUMMARY 4.3 *Validation of the MLE*

Objective: To evaluate the accuracy of the MLE as a function of the size of the dataset.

#1. Randomly select a value for the parameters of the considered model.

#2. Select the size of the dataset on which the estimation is performed and generate a corresponding number of observations, given the model and the choice of parameters.

#3. Perform the maximum likelihood estimation (Procedure Summary 4.2) on the previously generated dataset to estimate the parameters of the model.

#4. Evaluate the difference between the generator parameter and the estimated one as well as the difference between the probability pattern obtained with the generator parameter and the estimated one.

4.3.3 Simulated Transfer Data Analysis

In this subsection, Procedure Summary 4.3 is applied to transfer models in order to evaluate the accuracy of the MLE as a function of the size of the dataset.

Technical Aspects

i. The analysis included the following models: the GCM, the GCM-Lag and the three versions of the OGCM (OGCM-T, OGCM-L, OGCM-M).

ii. The studied categories were the 5-4 category set of Experiment I. All 16 learning and transfer items were considered.

iii. The accuracy of the estimation was tested on datasets characterized by the following lengths: 10, 40, 80, 120 and 160 blocks of 16 items. For each length, the procedure was iterated 100 times (the generator parameter was randomly selected each time).

iv. The gradient descent algorithm in the MLE was performed 10 times, each time starting from a random starting point.

Results

Figure 4.4 shows the relative error between the generator parameter and the estimated one, as a function of the size of the dataset, the model, and its parameters. The different transfer models are displayed on the columns, while their parameters are displayed on

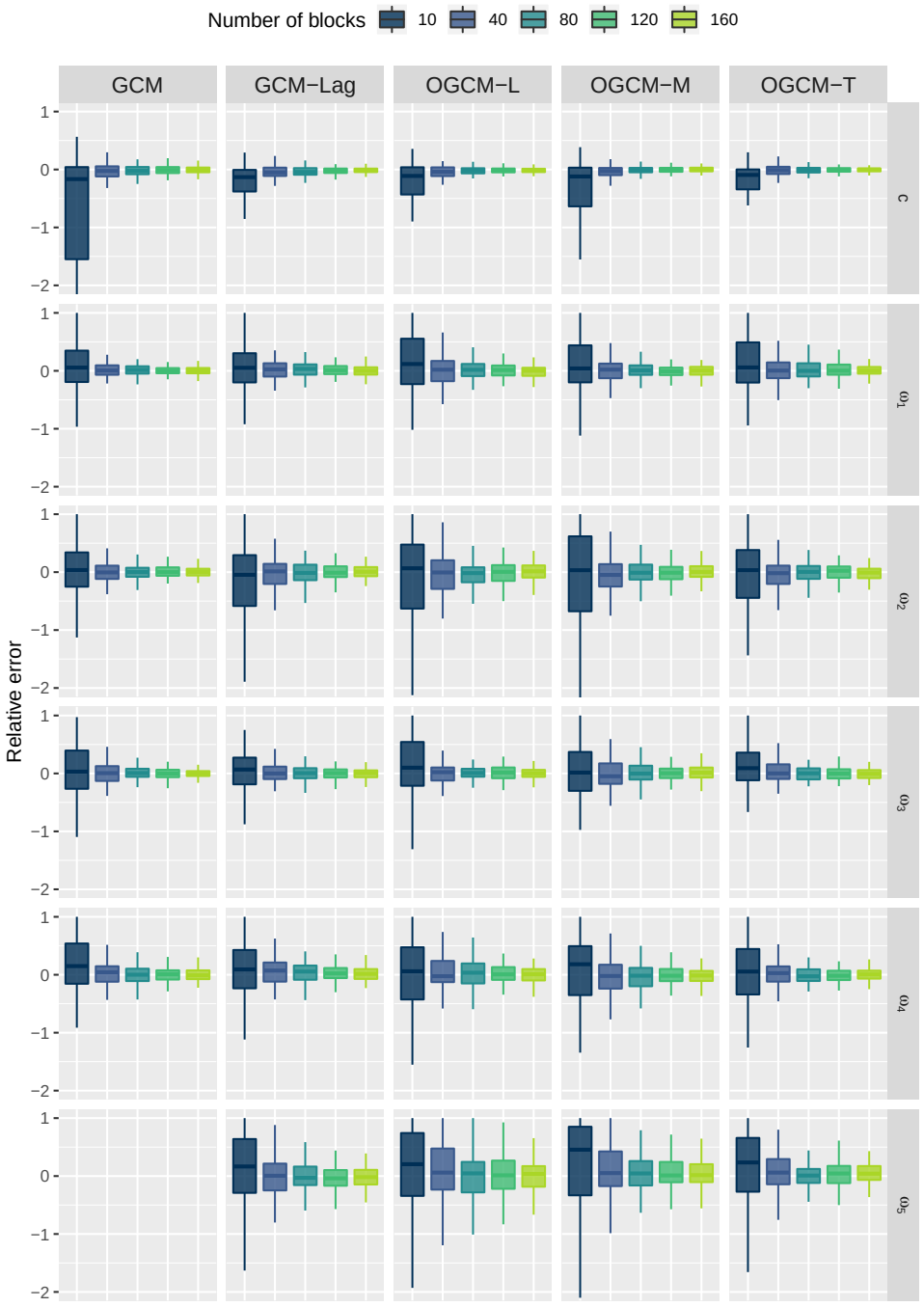


Figure 4.4 – *Box-plots representing the relative error of the maximum likelihood estimation on simulated transfer data, as a function of the size of the data, the studied transfer model, and its parameters. The relative error is defined as the difference between the generator parameter and the estimated one divided by the absolute value of the generator parameter. The box-plots were computed across 100 iterations and the generator parameter was randomly chosen at each iteration. The same items and categories of Experiment I were considered. The gradient descent algorithm in the MLE was performed 10 times.*

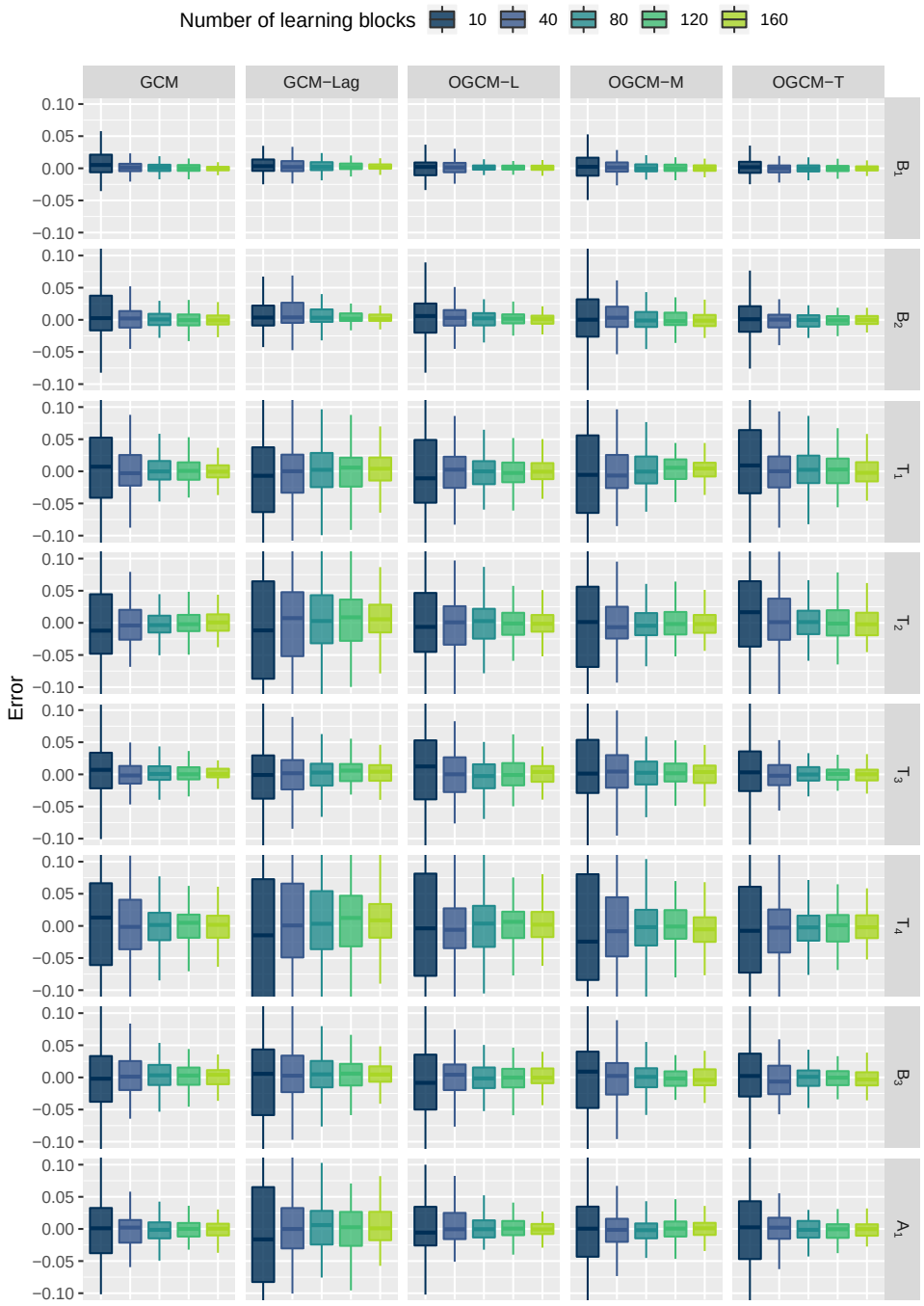


Figure 4.5 – *Box-plots representing the error between the probability pattern computed with the generator parameter and the one computed with the estimated parameter (on simulated transfer data), as a function of the size of the data, the transfer model, and the items. The box-plots were computed across 100 iterations and the generator parameter was randomly chosen at each iteration. The same items and categories of Experiment I were considered. The plot involves the first 8 items of Experiment I (the last 8 are shown in Figure 4.6). Items are denoted using the same notation as in Figure 2.1. The gradient descent algorithm in the MLE was performed 10 times.*

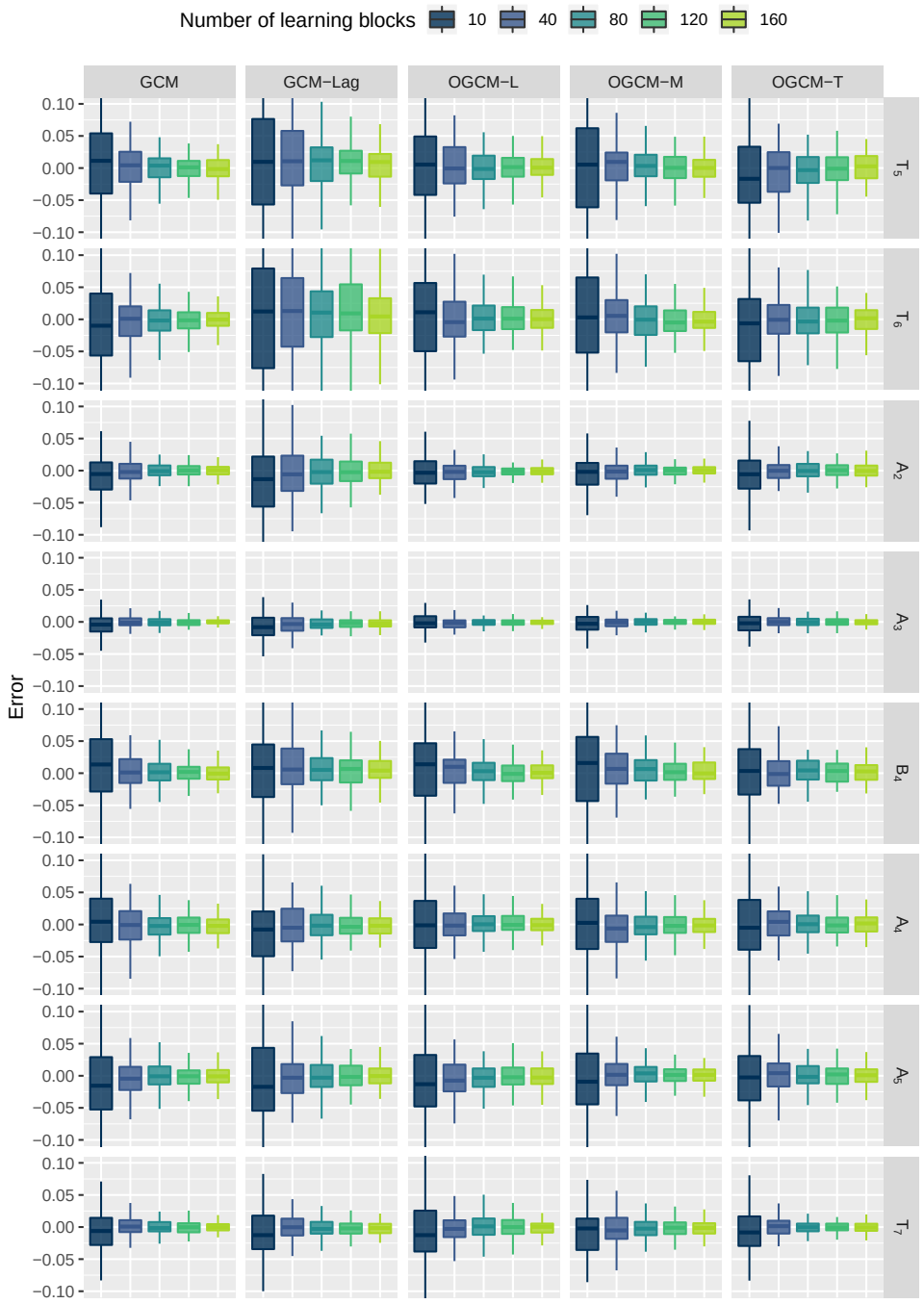


Figure 4.6 – *Box-plots representing the error between the probability pattern computed with the generator parameter and the one computed with the estimated parameter (on simulated transfer data), as a function of the size of the data, the transfer model, and the items. The box-plots were computed across 100 iterations and the generator parameter was randomly chosen at each iteration. The same items and categories of Experiment I were considered. The plot involves the last 8 items of Experiment I (the first 8 are shown in Figure 4.5). Items are denoted using the same notation as in Figure 2.1. The gradient descent algorithm in the MLE was performed 10 times.*

the rows. All relative errors (with some exceptions from the 10-blocks estimation) were centered around zero. Moreover, the variability of the relative error decreased as the number of blocks increased. There was no significant difference between the models. The parameter estimation seemed to be accurate when the size of the dataset is equal to or greater than 40 blocks.

Figure 4.5 and 4.6 show the error between the probability pattern computed with the generator parameter and the one computed with the estimated one, as a function of the size of the data, the model, and the item. The transfer models are displayed on the columns, while the items are displayed on the rows. The items are denoted using the same notation as in Figure 2.1. Again, the errors were overall centered around zero and their variability decreased as the number of blocks increased. Moreover, different items were not affected in the same way. For instance, the error on item B_1 was small, while the one on item T_4 was high. The error on learning items was generally smaller than the error on transfer items. There was no significant difference between the models. The estimation of the probability patterns seemed to be accurate when the size of the dataset was equal to or greater than 10-40 blocks.

4.4 Model Selection

A main approach to select the model that provides the best account of the data without overfitting them consists in using probabilistic statistical criteria such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC). These criteria prevent the risk of overfitting the data by penalizing the models on the basis of their number of parameters. The higher the number of parameters of a model, the higher its penalization. This penalization is fair when the number of parameters of a model corresponds to the dimension of its space of predictions. However, it is not clear whether the selected models meet this condition (see the analysis in Section 4.2). Therefore, cross-validation methods were preferred to probabilistic statistical criteria since their penalization does not involve the number of parameters of a models.

Two different cross-validation methods were used, depending on the nature of the models (i.e., learning or transfer). The k-fold cross-validation was used to compare transfer models, while the hold-out method (the simplest kind of cross-validation technique) was used to compare learning models. Since learning models are characterized by dependent observations (i.e., dependent random variables), no other cross-validation method was

adapted to this condition (further explanation will be given in Chapter 5). Conversely, the independence of the observations of transfer models enabled the use of more stable cross-validation methods. To recap, the choice to adopt two different cross-validation methods was motivated by the wish to use a more robust technique when it was licit. Since the k-fold cross-validation rely on the hold-out method, both methods are described in this chapter.

4.4.1 Hold-Out Method

The hold-out method is the simplest kind of cross-validation. It consists in separating the data in two sets: one that is used to estimate the parameters of the models and the other that is used to evaluate its predictions. Let us describe the hold-out method in detail. For the sake of simplicity, the description is given for transfer models.

Step #1

Let M be a transfer model and D a data-set. The first step consists in splitting the data-set into two sets: a training set D_L that is used to estimate the parameters of the model, and a testing set D_T that is used to quantify the difference between the predictions of the model and the real values. The use of two distinct sets prevents the risk that the model overfits the data.

Step #2

The second step consists in estimating the parameters of the model M by maximizing the likelihood $\mathcal{L}_M$ evaluated on the training set D_L (i.e., MLE). The maximum likelihood estimate $\hat{\theta}$ is found by means of Procedure Summary 4.2.

Step #3

Once the set of parameters have been estimated, they are used to determine the predictions of the model on the testing set. More precisely, for each stimulus of the testing set

$x^{(t)} \in D_T$ the probability of classifying the stimulus into category A (given M and $\hat{\theta}$) is computed:

$$\mathbb{P}_M^{\hat{\theta}}\left(A \,|\, x^{(t)}\right).$$

Step #4

The last step consists in computing the goodness-of-fit of the model on the testing set. We quantify the accuracy of the predictions in two ways: by considering either the Sum of Squared Deviations (SSD) or the likelihood. Let us detail both criteria.

Sum of Squared Deviations (SSD). The first criterion consists in summing across the testing set the squared difference between the prediction of the model and the real value (i.e., the participants' responses). Mathematically speaking, the sum of the squared deviations is given by:

$$E_{\text{SSD}} = \sum_{x^{(t)} \in D_T} \left(\mathbb{P}_M^{\hat{\theta}}\left(A \,|\, x^{(t)}\right) - z^{(t)}\right)^2, \tag{4.2}$$

where $z^{(t)}$ is the response given by the participant for the classification of the stimulus $z^{(t)}$. In mathematics, the SSD is usually called least-squares contrast.

Likelihood. The second criterion consists in using the likelihood of the model to determine how much the model is likely to generate the testing set when the parameter is fixed at $\hat{\theta}$. The higher the likelihood, the higher the probability that the testing set was produced by the model. The use of the opposite of the logarithm of the likelihood was preferred to the likelihood (see Remark 4.1). Thus, the evaluation of the model using the likelihood is given by:

$$E_{\mathcal{L}} = -\log \mathcal{L}_M(D_T\,;\,\hat{\theta}). \tag{4.3}$$

The two criteria are largely used both in mathematics and psychology. In mathematics, the use of the likelihood is generally preferred, especially when the parameter estimation is performed using the MLE. Conversely, in psychology, the use of the sum squared deviations is more popular than the likelihood. Here a selection of studies using the SSD: [CG19; NKM92; Nos+94; Nos+18; Pal99]. Both criteria were adopted to simultaneously provide a more robust evaluation and allow a continuity with previous studies.

PROCEDURE SUMMARY 4.4 *Hold-Out Method*

Objective: To quantify how accurately a model reproduced a data sample.

#1. Split the data sample into training and testing sets.

#2. Perform the maximum likelihood estimation (MLE) on the training set to estimate the parameters of the model (Procedure Summary 4.2).

#3. Use the previously estimated parameters to compute the model predictions on the testing set.

#4. Evaluate the predictions of the model on the testing set using both the sum squared deviations (SSD) and the likelihood.

4.4.2 k-Fold Cross-Validation

The k-fold cross-validation represents one way to improve over the hold-out method. It consists in splitting the data in k sets and applying the hold-out method k times. Each time, a different set is used as the testing set and the remaining $k-1$ sets are used as the training set. Since the k-fold cross-validation is reasonably straightforward once the hold-out method is understood, it is only described by means of the framework below (the description is given for transfer models).

PROCEDURE SUMMARY 4.5 *k-Fold Cross-Validation*

Objective: To quantify how accurately a model reproduces a data sample. This cross-validation method is generally more stable than the hold-out method.

Let M be a transfer model and D a data-set, the k-fold cross-validation is composed of the following steps:

#1. Split the data-set D in k sets. Consider one of them as the testing set and the $k-1$ remaining sets as the training set.

#2. Apply the hold-out method (Procedure Summary 4.4) to the considered training and testing sets. Store the evaluation of the goodness-of-fit of the model expressed in terms of SSD or likelihood.

#3. Repeat the process until all k sets serve as the testing set. Since the data-set has been separated into k sets, the hold-out method is repeated k times (each time using a different set as the testing test).

#4. Average the k recorded evaluations resulting from the repetitions of the hold-out method. The averaged result represents the final evaluation of the goodness-of-fit of the model.

4.4.3 Validation of the k-Fold Cross-Validation

Since the description of the k-fold cross-validation is completed, we could be tempted to apply the technique to experimental data directly. Let us imagine this situation for a moment. Let us say that the result of the application of the k-fold cross-validation to a data sample shows that the best fit is provided by a certain model. Does this mean that the experimental data have been generated by the model that best fits them? Or, at least, does this mean that (among the considered models) the model with the best evaluation is the one having the highest probability to be the generator model? At this stage, we are not able to answer these questions.

The aim of this subsection is to investigate whether the transfer models are identifiable via the k-fold cross-validation. If models are not identifiable, even the more robust method would fail at identifying them. In what follows, we describe the procedure we used to determine whether the transfer models were identifiable via the k-fold cross-validation. Let $M \in \mathfrak{M}$ be a transfer model and θ_M its parameter.

Step #1

The first step consists in generating a set of data, according to the model M and the value of its parameters θ^M. In more detail, given the sequence of items $x^{(1)}, \ldots, x^{(n)}$ and its corresponding sequence of feedback $v^{(1)}, \ldots, v^{(n)}$, a sequence of responses $z^{(1)}, \ldots, z^{(n)}$ is

generated. This sequence of responses satisfy the following condition: $z^{(i)}$ is a realization of

$$Z^{(i)} \sim \mathcal{B}\left(\mathbb{P}_M^{\theta_M}\left(A \,|\, x^{(i)}\right)\right).$$

The sequence of responses generated with the model M is denoted by D_M.

Step #2

Once the set of observations D_M have been generated, the second step consists in determining the model (among the models of $\mathfrak{M}$) that best fits the data sample D_M using the k-fold cross-validation. In detail, Procedure Summary 4.5 is applied to data sample D_M for each $m \in \mathfrak{M}$. By applying this procedure, the evaluations of the models on D_M using both the SSD and the likelihood are computed for each $m \in \mathfrak{M}$. The evaluations are denoted by $E_{\text{SSD}}^{m,M}$ and $E_{\mathcal{L}}^{m,M}$.

Step #3

The last step consists in comparing the evaluations obtained in the previous step. Our hope is that the k-fold cross-validation detected the model with which the data sample have been generated. In other words, we hope that the model that generated the observations is the model with the smallest evaluation:

$$\arg\min_{m \in \mathfrak{M}} E_{\text{SSD}}^{m,M} \overset{?}{\in} M \qquad \text{or} \qquad \arg\min_{m \in \mathfrak{M}} E_{\mathcal{L}}^{m,M} \overset{?}{\in} M$$

These three steps are iterated multiple times to give a statistical significance to the study.

PROCEDURE SUMMARY 4.6 *Validation of the k-Fold Cross-Validation*

Objective: To verify whether the k-fold cross-validation (Procedure Summary 4.5) is able to detect the model (among a set of models) that generated the simulated data.

#1. Select a model from a set of models and chose a value for its parameters. Generate a data sample according to the selected model and value of parameter.

#2. Apply the k-fold cross-validation (Procedure Summary 4.5) to the previously generated data sample for each model of the set of models.

#3. Compare the evaluations of the models. Determine whether the k-fold cross-validation detected the generative model.

4.4.4 Simulated Transfer Data Analysis

In this subsection, Procedure Summary 4.6 is applied to transfer models to determine whether the k-fold cross validation is able to detect the model underlying the simulated observations.

Technical Aspects

i. The set of models $\mathfrak{M}$ included the following models: the GCM, the GCM-Lag, and the three versions of the OGCM (OGCM-T, OGCM-L, and OGCM-M).

ii. Since the aim was to validate the results of the k-fold cross-validation applied to the transfer phase of Experiment I, the same sequence of stimuli as in the transfer phase of Experiment I was considered. The number of blocks of the simulated dataset was equal to 5 blocks $\times$ 43 participants.

iii. A 5-fold cross-validation was used. We anticipate that the 5-fold is the type of k-fold that will be applied to Experiment I.

iv. The parameters used to simulate the responses are the average parameters resulting from the MLE during the application of the 5-fold to the transfer phase of Experiment I.

v. The procedure was iterated 100 times. Thus, for each model, 100 sequences of responses associated with the sequence of stimuli of the transfer phase of Experiment I were generated. A 5-fold was applied to each sequence of responses.

vi. At each MLE, the gradient descent algorithm was performed 10 times, each time starting from a random starting point.

Figure 4.7 – *Validation of the k-fold cross-validation (Procedure Summary 4.6) on simulated data having the same features as Experiment I. The graph shows the number (and percentage) of times that the considered models obtained the lowest evaluation (by using the SSD or the likelihood), giving a specific simulated data sample. The procedure was iterated 100 times. We adopted a 5-fold cross-validation. Models are evaluated with both the SSD and the likelihood.*

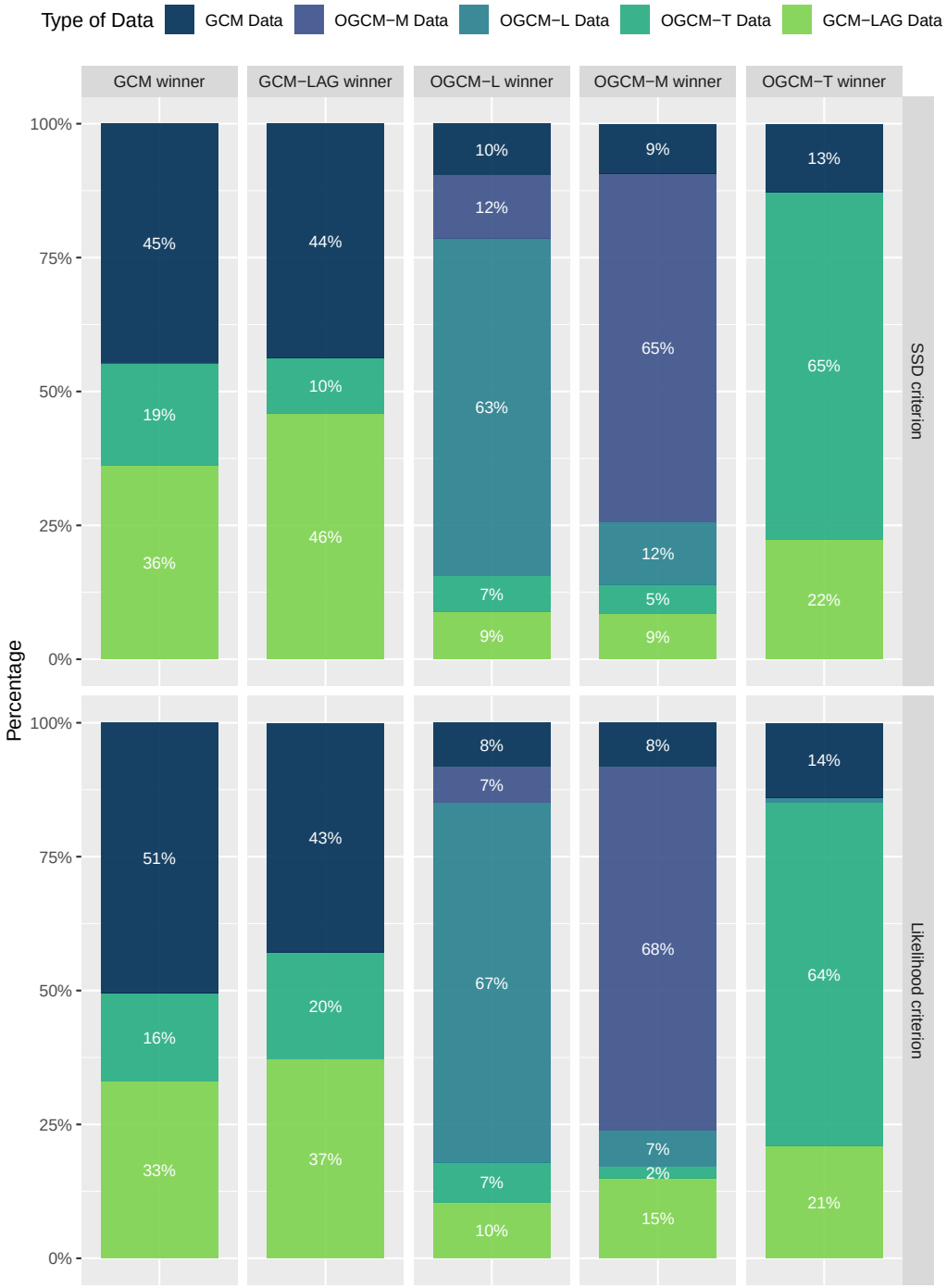


Figure 4.8 – *Validation of the k-fold cross-validation (Procedure Summary 4.6) on simulated data having the same features as Experiment I. The graph shows the percentage of times that the simulated data were actually generated by the model that has the lowest evaluation (by means of the SSD or the likelihood). The procedure was iterated 100 times. We adopted a 5-fold cross-validation. Models are evaluated with both the SSD and the likelihood.*

Results

The results of the application of Procedure Summary 4.6 to transfer models on simulated data having the same features as Experiment I are shown in Figure 4.7 and 4.8. The first graph shows the number (and percentage) of times that the considered models obtained the lowest evaluation (by using the SSD or the likelihood), giving a specific simulated data sample. The second graph adopts a different prospective and shows the percentage of times that the simulated data were actually generated by the model that has the lowest evaluation (by means of the SSD or the likelihood). Although the two graphs give a complementary insight on the identifiability of the models via the 5-fold method, the second graph is more useful to interpret the results on experimental data.

The criterion with which the models were evaluated did not remarkably affect the results. The most identifiable models were the OGCM-L and the OGCM-M (their data were correctly identified 84-91% of time), followed by the OGCM-T (their data were correctly identified 61-64% of time), the GCM (their data were correctly identified 42-49% of time), and the GCM-Lag (their data were correctly identified 13-22% of time). Moreover, the OGCM-L data were misidentified as OGCM-M data 9-15% of time (and conversely). The OGCM-T data was most often misidentified as GCM data, the GCM data was most often misidentified as GCM-Lag data, and the GCM-Lag data was most often misidentified as GCM data.

Let us now analyze Figure 4.8. When the model with the lowest evaluation was either the OGCM-L, or the OGCM-M, or the OGCM-T, then it was the generative model with a probability of 63-68%. Conversely, when the model with the lowest evaluation was either the GCM or the GCM-Lag, then it was the generative model with a probability of 37-51%.

Moreover, when the model with the lowest evaluation was either OGCM-L or OGCM-M, then the generative model was a model that accounts for the order during the learning phase 75-77% of the time. When the model with the lowest evaluation was the OGCM-T, then the generative model was a model that accounts for the order during the transfer phase 85-87% of the time. Finally, when the model with the lowest evaluation was the GCM-Lag, then the generative model was a model that accounts for the order during the transfer phase with a probability of 45-50%.

	SSD	-log$\mathcal{L}$	c	ω_1	ω_2	ω_3	ω_4	ω_O
GCM	93.1	293.8	7.6	0.22	0.2	0.37	0.21	-
GCM-Lag	93.1	293.8	7.6	0.21	0.19	0.37	0.2	0.03
OGCM-L	92	290.7	7.4	0.17	0.14	0.36	0.15	0.17
OGCM-M	**91.4**	**288.9**	7.3	0.16	0.13	0.36	0.14	0.21
OGCM-T	92.9	293.5	7.7	0.2	0.19	0.37	0.19	0.06

Table 4.2 – *Goodness-of-fit of the transfer models and average estimated parameters resulting from the application of a 5-fold (Procedure Summary 4.5) to the transfer phase of Experiment I. Model were evaluated by means of both the SSD the likelihood.*

4.5 Experimental Transfer Data Analysis

The aim of this chapter is to apply the k-fold (more specifically, the 5-fold) to the transfer models on the transfer phase of Experiment I. The 5-fold was applied three times: the first time to all participants, the second time only to participants in the rule-based order, and the third time only to participants in the similarity-based order.

Technical Aspects

i. The analysis included the following models: the GCM, the GCM-Lag, and the three versions of the OGCM (OGCM-T, OGCM-L, and OGCM-M).

ii. A 5-fold cross-validation was used.

iii. At each MLE, the gradient descent algorithm was performed 10 times, each time starting from a random starting point.

Results

Application of the 5-fold to all participants. The goodness-of-fit of the transfer models to the transfer phase of Experiment I are shown in Table 4.2. The OGCM-M is the model with the lowest evaluation with both criteria (SSD and likelihood). The analysis performed in the previous section ensures that the OGCM-M is the generative model with

a probability of 65-68%. Moreover, it ensures that the generative model is a model that accounts for the order received during the learning phase (i.e., OGCM-L or OGCM-M) with a probability of 75-77%. Table 4.2 also shows the estimated parameters of the models averaged on the 5 hold-outs that compose the 5-fold. The attention-weight that regulates the ordinal dimension was not negligible in both the OGCM-L and OGCM-M ($\omega_O = 0.17$ for the OGCM-L and $\omega_O = 0.21$ for the OGCM-M). This shows that the integration of the stimuli order received during the learning phase allowed both models to better reproduce performance. Conversely, in both the GCM-Lag and the OGCM-T, the attention-weight that regulates the ordinal dimension was negligible ($\omega_O = 0.03$ for the GCM-Lag and $\omega_O = 0.06$ for the OGCM-T), showing that the integration of the stimuli order received during the transfer phase did not allow both models to better describe the data.

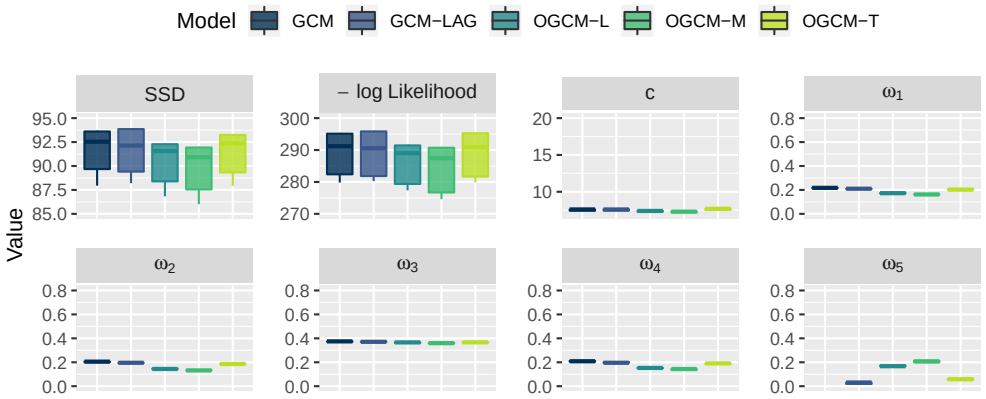
Figure 4.9 (on the top) shows the variation across the 5 hold-outs of both the evaluation criteria and the estimated parameters, as a function of the model. The values of both the SSD and likelihood evaluations are similar across models. Similarly, the values of the estimated parameters are similar across models, except for the attention-weight parameter that regulates the ordinal dimension. Again, its value was negligible for the GCM-Lag and the OGCM-T. Finally, ω_3 was the attention-weight parameter with the highest value. This is not surprising since the third dimension (which is associated to the color, see Figure 2.1) is the feature dimension that allows participants to reach the highest proportion of correct responses.

Finally, Figure 4.10a and 4.10b (first column on both graphs) shows the predictions of the models as a function of the items. The learning items are displayed in Figure 4.10a (as rows), while transfer items are displayed in Figure 4.10b (as rows). The participants' transfer performance are indicated with an x-mark (they were averaged across participants and blocks). The graph shows that all models achieved very good quantitative predictions on both learning and transfer items. The fact that the goodness-of-fit of the models was similar overall and that all models provided good predictions suggests that the benefit to integrate the order received during learning is modest. Further investigation are needed to shed light on the role of integrating stimuli manipulation on models.

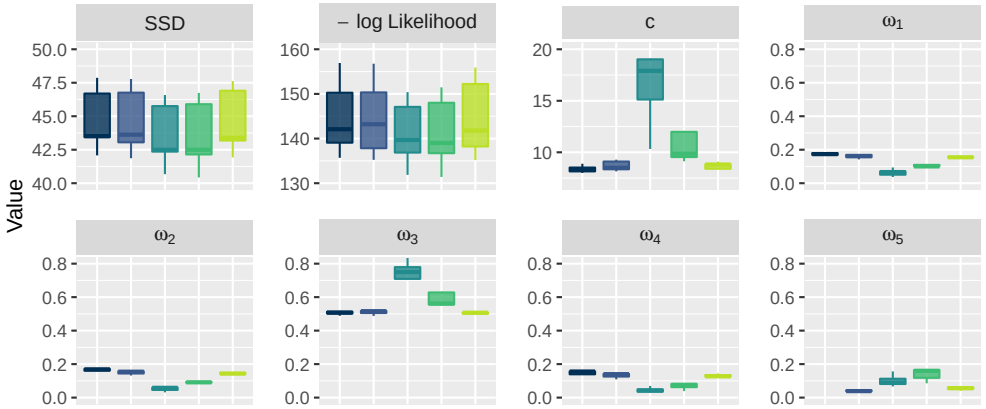
Application of the 5-fold to participants in the rule-based and similarity-based orders. Figure 4.9 (in the middle and on the bottom) shows the variation across the 5 hold-outs of both the evaluation criteria and the estimated parameters, when the 5-fold

was only applied to either participants in the rule-based order (in the middle) or participants in the similarity-based order (on the bottom). The estimated attention-weight parameter of the third dimension was higher for participants in the rule-based order than for participants in the similarity-based order (especially in the OGCM-L). This means that the models (in particular the OGCM-L and OGCM-M) detected that participants in the rule-based order relied more on the third dimension than participants in the similarity-based order to classify stimuli. In other words, they identified that participants in the rule-based order adopted a rule-based strategy more often than participants in the similarity-based order. Moreover, they detected that the performance of participants in the rule-based order was slightly higher than the performance of participants in the similarity-based order (the values of the estimated sensitive parameter were higher in the graph in the middle than the graph on the bottom).

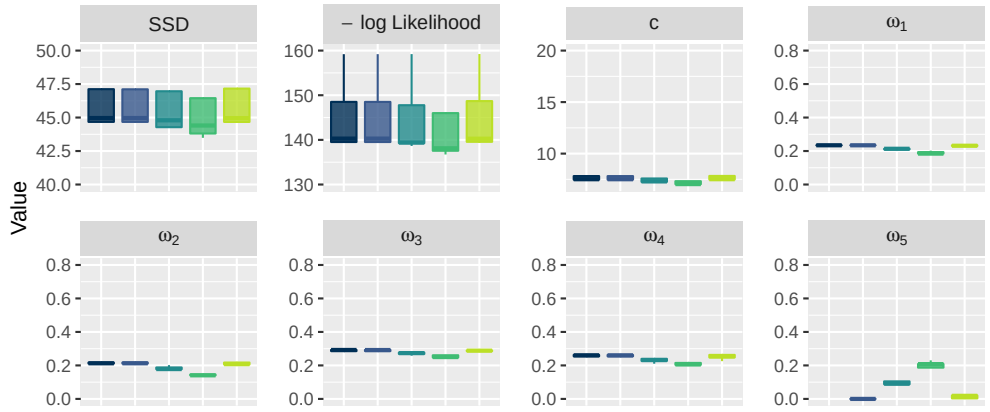
Figure 4.10a and 4.10b (second and third column) shows the predictions of the models as a function of the items, when the 5-fold was only applied to either participants in the rule-based order (second column) or participants in the similarity-based order (third column). Again, the learning items are displayed in Figure 4.10a (as rows), while transfer items are displayed in Figure 4.10b (as rows). The participants' transfer performance are indicated with an x-mark (they were averaged across participants and blocks). The OGCM-L and OGCM-M seems to be the models that best adapted their predictions to the stimuli manipulation received by participants. However, all models achieved very good quantitative predictions on both learning and transfer items.



(a) *All participants.*



(b) *Participants following a rule-based study.*



(c) *Participants following a similarity-based study.*

Figure 4.9 – *Box-plots representing the values of the SSD evaluation, the likelihood evaluation, and the estimated parameters during the application of the 5-fold cross-validation on the transfer phase of Experiment I. The 5-fold was applied three times: to all participants (on the top), to participants in the rule-based order (in the middle), and to participants in the similarity-based order (on the bottom).*

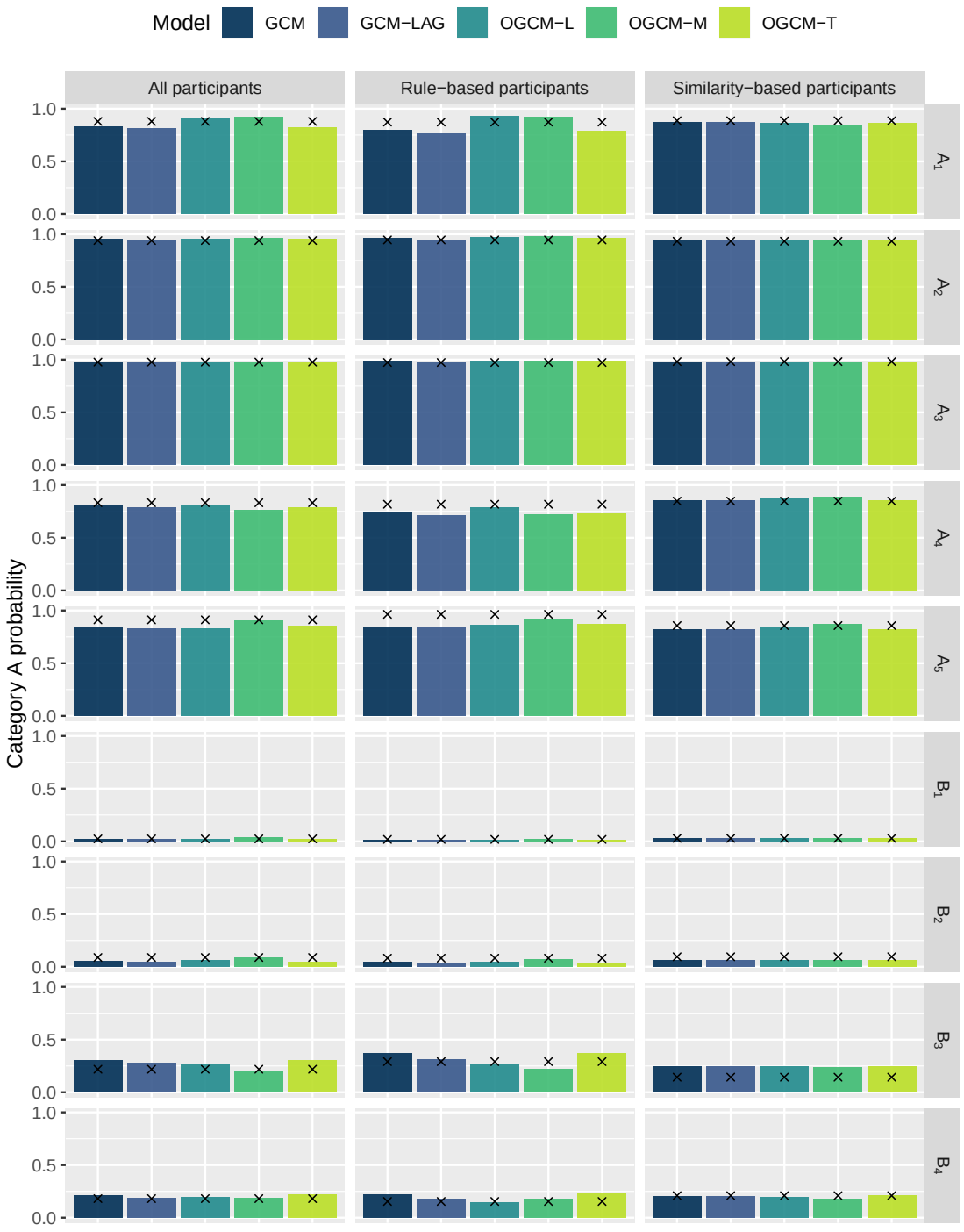


Figure 4.10a – *Predictions of the transfer models on the transfer phase of Experiment I, as a function of the items. The 5-fold was applied three times: to all participants, to participants in the rule-based order, and to participants in the similarity-based order. Only learning items are displayed. The participants' transfer performance are indicated with an x-mark. Both the predictions of the models and the participants' performance were averaged (the first across the 5 hold-outs and the second across participants).*

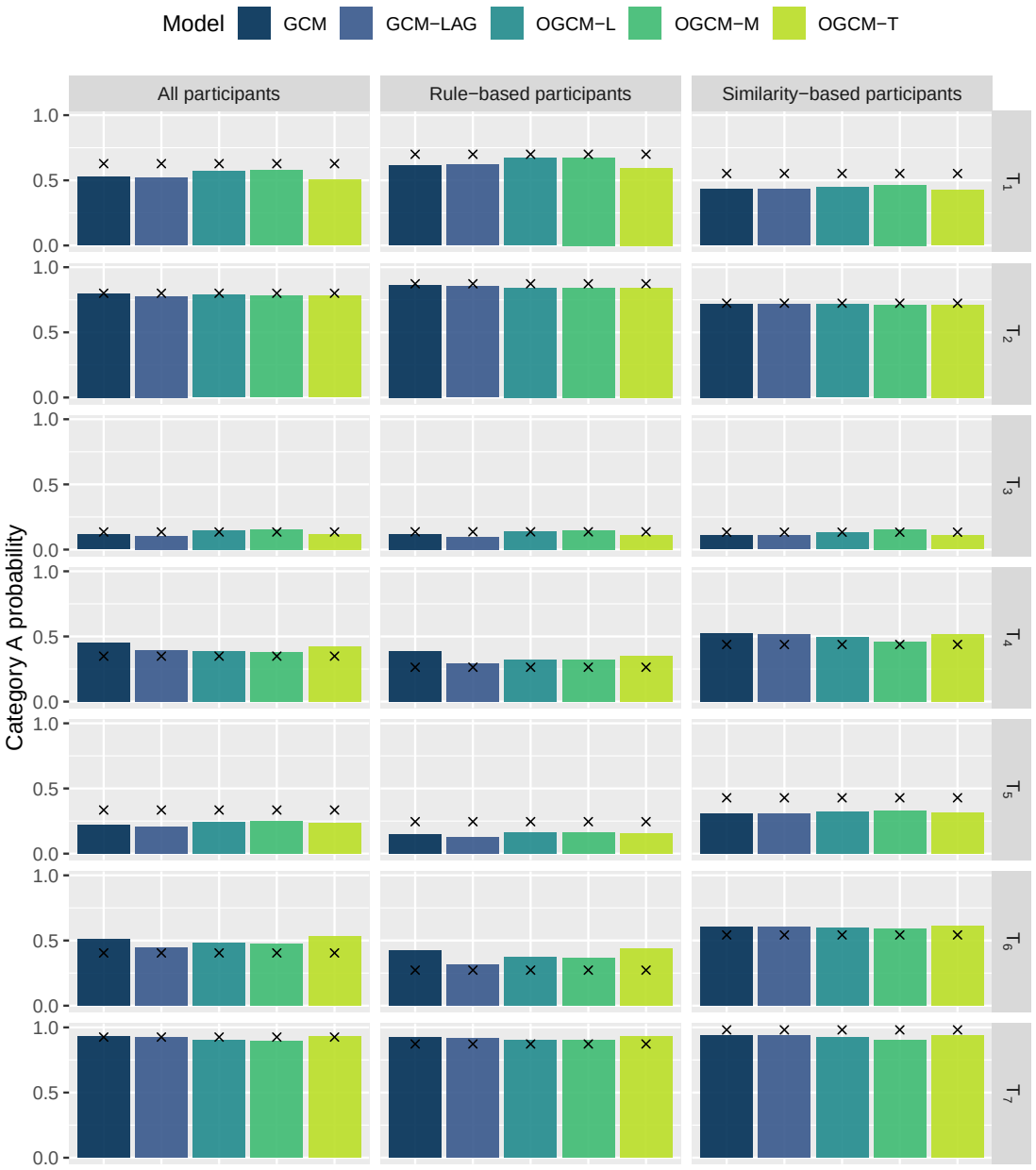


Figure 4.10b – *Predictions of the transfer models on the transfer phase of Experiment I, as a function of the items. The 5-fold was applied three times: to all participants, to participants in the rule-based order, and to participants in the similarity-based order. Only transfer items are displayed. The participants' transfer performance are indicated with an x-mark. Both the predictions of the models and the participants' performance were averaged (the first across the 5 hold-outs and the second across participants)*

TO SUM UP *Advanced Inference Method and Application to Transfer Models*

The aim of this chapter is twofold. On one hand, to develop a general and robust inference method to compare models. On the other hand, to apply this inference method to transfer models to determine the model that best describes the transfer phase of Experiment I.

Visual Representation of Models

Because of the complexity of the models, we estimated that a preliminary visualization of the models was warranted. This analysis investigated the spreading of the predictions of the transfer models using the Principal Component Analysis (PCA). The static variant of the learning models were also included. The results showed that the transfer models occupied a limited region of the prediction space, while the static variant of the learning models covers the prediction space entirely. The GCM-Lag was the transfer model with the greatest spreading, while the other transfer models had predictions with a similar variability.

Parameter Estimation

The parameters of the models were estimated by means of the Maximum Likelihood Estimation (MLE) and the MLE was implemented using the gradient descent algorithm. Since the unclear identifiability of the transfer models did not guarantee the consistency of the MLE, the accuracy of the estimation was evaluated on simulated data. The analysis showed that the parameter estimation seemed to be accurate from 40 blocks, while the estimation of the classification probability seemed to be accurate from 10-40 blocks.

Model Selection

Because of the unclear correspondence between number of parameters and dimension of the prediction space, cross-validation methods were preferred to probabilistic statistical criteria such as BIC or AIC. The k- fold was used to compare transfer models, while the hold-out was used to compare learning models. The evaluation

of the model was performed using both the Sum Squared Deviations (SSD) and the likelihood. A preliminary validation of the identifiability of the transfer models via the k-fold was necessary to allow us to evaluate the results on experimental data. The results showed that when the model with the lowest evaluation was either the OGCM-L, or the OGCM-M, or the OGCM-T, then it was the generative model with a probability of 63-68%. Conversely, when the model with the lowest evaluation was either the GCM or the GCM-Lag, then it was the generative model with a probability of 37-51%.

Experimental Transfer Data Analysis

The 5-fold cross validation was applied to transfer models to determine the model that best describes the transfer phase of Experiment I. The transfer model that best fits the transfer phase of Experiment I was the OGCM-M, while the second model was the OGCM-L. The estimated attention-weight parameter that regulates the ordinal dimension was not negligible in both the OGCM-M and OGCM-L, showing that the information provided by the ordinal dimension was relevant for the classification. Conversely, the estimated ordinal attention-weight parameter was negligible in both the OGCM-T and GCM-Lag. However, the goodness-of-fit of the models was similar overall and all models provided good predictions. Finally, the separate application of the 5-fold to participants in the rule-based and similarity-based orders showed that *i)* the models (in particular the OGCM-L and OGCM-M) detected that the majority of the participants in the rule-based order adopted a rule-based strategy, and *ii)* the OGCM-L and OGCM-M seemed to be the models that best adapted their predictions to the stimuli manipulation.

5

Application of the Advanced Inference Method to Learning Models

Contents

Transfer or learning performance of ALCOVE and Component-Cue have been evaluated in the literature [CMB93; GB88a; GB88b; GBH89; Kru92; LN02; Nos+94; Pal99; RR04; SE15]. However, only a limited number of studies have compared the performance of ALCOVE and Component-Cue on a same classification task [NKM92]. For example, Nosofsky [NKM92] evaluated the context model, ALCOVE, and Component-Cue on two experiments: the first that partially replicated and extended the probabilistic classification learning paradigm of Gluck and Bower [GB88a], and the second that extended the classification learning paradigm of Medin and Schaffer [MS78]. Nosofsky found that only the exemplar-based network (i.e., ALCOVE) achieved good quantitative predictions of the learning and transfer data in both experiments.

Goals

In view of the above, the first goal of the present chapter is to further investigate the comparison between ALCOVE and Component-Cue. The inference method developed in the previous chapter (see Chapter 4) is here applied to learning models to determine the one that best describes both Experiment I and Experiment II data-sets.

The second goal is to investigate whether the distinct network architecture of ALCOVE and Component-Cue (indeed, the nodes of the networks code items in the case of ALCOVE and features in the case of Component-Cue) is related to the within-category presentation order (i.e., rule-based and similarity-based). Again, although ALCOVE and Component-Cue integrate the same error-driven mechanism, they express two different learning strategies: a similarity-based strategy for ALCOVE and an induction strategy (i.e., a rule-based strategy) for Component-Cue. Therefore, one plausible hypothesis is that ALCOVE would better reproduce the performance of participants adopting a similarity-based strategy (as compared to those adopting a rule-based strategy), while Component-Cue would better reproduce the performance of participants adopting a rule-based strategy (as compared to those adopting a similarity-based strategy). Moreover, since it has been shown that participants in the rule-based order usually exhibit generalization patterns consistent with rule-based retrieval (i.e., they adopted a rule-based strategy) [MF16], our intuition is that Component-Cue would be more adapted to participants in the rule-based order. Conversely, ALCOVE would be more adapted to participants in the similarity-based order.

Outline of this chapter

Firstly, we provide a visual representation of both the predictions and learning curves of the learning models. Secondly, we evaluate the consistency of the Maximum Likelihood Estimation (MLE) as a function of the size of the dataset when learning models are considered. Then, we recall the cross-validation technique that was adopted to compare learning models. Finally, we apply this cross-validation technique to determine the learning model that best describes the data in Experiment I and II, and we investigate the relation between the studied models and the within-category order.

Note: The mathematical difference between learning and transfer models implied small modifications in both the inference method and its formalization. Firstly, since learning models can also account for the learning dynamics, the predictions of the model during the

learning phase were also analyzed. Secondly, since the description of the inference method was given for transfer models (see Chapter 4), few modifications were implemented.

5.1 Visual Representation of Models

The non-independence of the predictions of the learning models (due to the error-driven mechanism) adds complexity to the already intricate context (unclear correspondence between the number of parameters and the dimension of the prediction space, large number of parameters, etc.). Therefore, we considered that a preliminary visualization of the learning models would have helped to better understand their behavior.

The analysis is organized in two parts. The first part aims to analyze the predictions of the models after a period of training (i.e., the predictions during the transfer phase), while the second part aims to analyze the learning curves of the models (i.e., the predictions during the learning phase).

5.1.1 Simulated Transfer Data Analysis

In this subsection, Procedure Summary 4.1 is adapted to learning models and applied to them. Since the predictions of the learning models depend on the stimuli and feedback used to train them, this dependency has to appear in the computation of the probability patterns. Let us say that the model has been trained on n stimuli. Therefore, the function g_M in Equation 4.1 is modified as follows:

$$\begin{aligned} g_M : \Theta_M &\longrightarrow [0,1]^N \\ \theta_M &\longmapsto P_M^{\theta_M} = \left(\mathbb{P}_M^{\theta_M}\left(A \,|\, \xi_1, \mathcal{H}_n\right), \ldots, \mathbb{P}_M^{\theta_M}\left(A \,|\, \xi_N, \mathcal{H}_n\right) \right), \end{aligned}$$

where $\mathcal{H}_n$ represents the n stimuli and feedback on which the model has been trained (i.e., the history of the process as defined in Section 3.1). In other words, the predictions of the models during the transfer phase depend on the n stimuli and feedback that the model received during the training.

Technical Aspects

i. The analysis included the following models: Component-Cue$^{\mathrm{L}}$, Component-Cue$^{\mathrm{E}}$, ALCOVE$^{\mathrm{L}}$ and ALCOVE$^{\mathrm{E}}$.

ii. The studied categories were the 5-4 category set of Experiment I (see Figure 2.1). Moreover, the probability pattern included the classification probability of all training and transfer items of Experiment I ($N = 16$).

iii. The models were previously trained on 10 random blocks of 9 (learning) items, i.e. $n = 9$ stimuli $\times 10$ blocks $= 90$ stimuli.

iv. A total of 20 000 sets of parameters per model was considered ($l = 20\,000$). Each set was randomly chosen among those respecting the constraints of the models.

Results

Figure 5.1a and 5.1b shows three different planes of the PCA resulting from the application of the modified Procedure Summary 4.1 to learning models. The position of a set of relevant probability patterns is highlighted in both figures to facilitate the interpretation of the graphs. Since the studied categories were the same as those considered in the visual representation of the transfer models (see Section 4.2), the same set of relevant patterns was used. A visual description of the relevant patterns is given in Figure 5.1a (on the bottom), while a verbal description can be found in Subsection 4.2.2.

The analysis of the PCA planes showed that the learning models are nested. The following hierarchy is observed (starting from the model with the smallest spreading): Component-Cue$^{\mathrm{L}}$, Component-Cue$^{\mathrm{E}}$, ALCOVE$^{\mathrm{L}}$ and ALCOVE$^{\mathrm{E}}$. The ALCOVE models (i.e., exponential and linear versions) attained a larger range of probability patterns as compared to the Component-Cue models (i.e., exponential and linear versions). This is probably due to the fact that ALCOVE has a greater number of parameters as compared to the Component-Cue.

Moreover, the linear versions attained a more limited range of probability patterns as compared to the exponential versions. Although the linear and exponential versions have the same number of parameters, the exponential version is richer in variability as compared to the linear version. This is due to the distinct role of the parameters associated to each version (see Remark 3.8). This argument goes in the direction of the

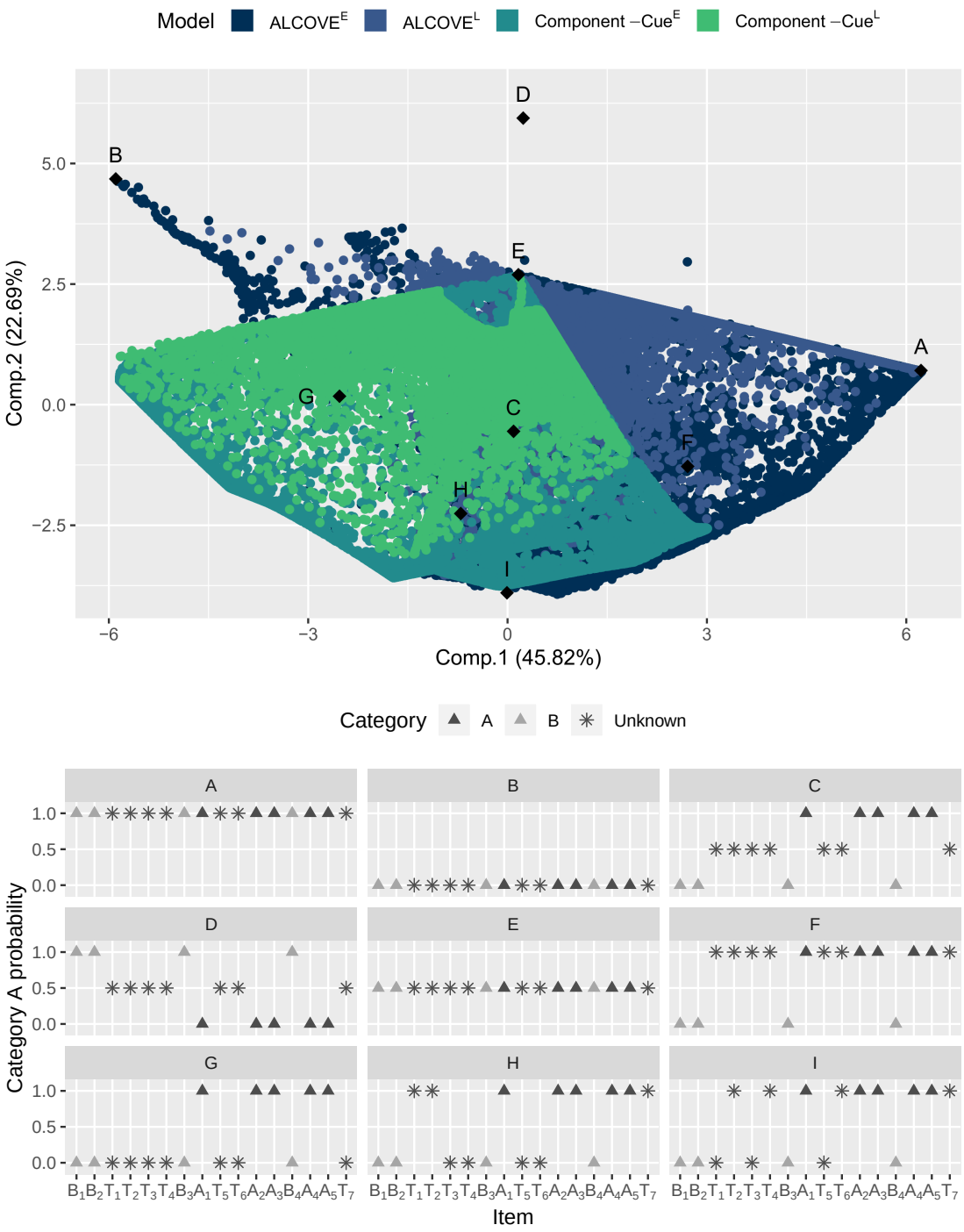


Figure 5.1a – *Principal Component Analysis (PCA) planes resulting from the application of the modified Procedure Summary 4.1 to learning models. The patterns refer to the 5-4 category set of Experiment I (transfer items included). On the top, the projection of the probability patterns on the first and second components. The models were trained on 10 blocks of 9 stimuli. A total of 20 000 patterns per model were considered. On the bottom, some relevant probability patterns.*

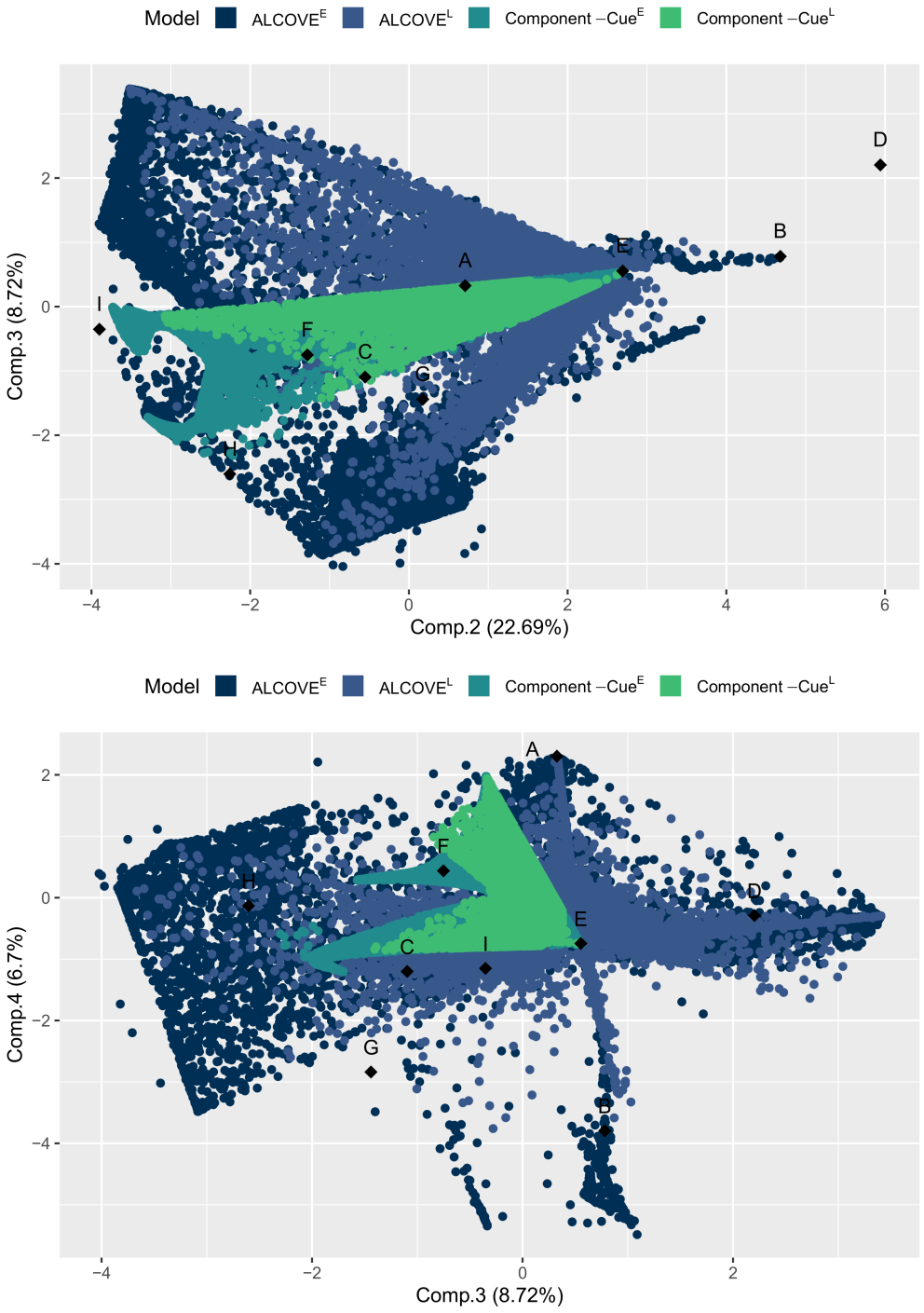


Figure 5.1b – *Principal Component Analysis (PCA) planes resulting from the application of the modified Procedure Summary 4.1 to learning models. The patterns refer to the 5-4 category set of Experiment I (transfer items included). On the top, the projection of the probability patterns on the second and third components. On the bottom, the projection on the third and forth components. The models were trained on 10 blocks of 9 stimuli. A total of 20 000 patterns per model were considered.*

non-correspondence between number of parameters and dimension of the prediction space (Section 4.2).

Learning models were unable to reach patterns in which learning items are classified into their opposite category (i.e., pattern D). This in probably due to the error-driven mechanism. Indeed, the updating mechanism is tailored to increase the probability to correctly classify the current item (see Subsection 3.3.1 and 3.3.2). Therefore, although the updating could decrease the probability to correctly classify other items, it necessarily benefits the current item. This prevents the model to reach totally incorrect patterns. The same analysis conducted on the categories of Experiment II found similar results.

5.1.2 Simulated Learning Data Analysis

In this subsection, we analyze the predictions of the learning models during the learning phase, which are called learning curves. More specifically, the aim is to determine how the parameters of the models influence the learning curves. For this purpose, both a high and a low values of each parameter are considered to detect changes in the learning curves.

Technical Aspects

i. The analysis included the following models: Component-Cue$^{\mathrm{L}}$, Component-Cue$^{\mathrm{E}}$, ALCOVE$^{\mathrm{L}}$ and ALCOVE$^{\mathrm{E}}$.

ii. The studied categories were the 5-4 category set of Experiment I (see Figure 2.1). Since the investigation involved the learning phase, only the 9 learning items were considered.

iii. The models were trained on 50 random blocks of 9 learning items, i.e. $n = 50$ blocks $\times$ 9 learning items $= 450$ stimuli.

iv. The values of the parameters used to compute the learning curves of Figure 5.2a and 5.2b are shown in Table 5.1. We recall that the parameters of ALCOVE^{E} are c, ϕ, λ_ω and λ_w; those of ALCOVE^{L} are c, b, λ_ω and λ_w; those of Component-Cue^{E} are λ_w and ϕ; and those of Component-Cue^{L} are λ_w and b.

	High					Low				
	c	λ_ω	λ_w	b	ϕ	c	λ_ω	λ_w	b	ϕ
Parameters λ_ω, λ_w										
ALCOVE^{E}	5	**.5**	**1**	-	5	5	**.005**	**.01**	-	5
ALCOVE^{L}	5	**.5**	**1**	1	-	5	**.005**	**.01**	1	-
Component-Cue^{E}	-	-	**.5**	-	5	-	-	**.005**	-	5
Component-Cue^{L}	-	-	**.5**	1	-	-	-	**.005**	1	-
Parameter c										
ALCOVE^{E}	**10**	.005	.01	-	1	**.5**	.005	.01	-	1
ALCOVE^{L}	**10**	.005	.01	1	-	**.5**	.005	.01	1	-
Parameter ϕ										
ALCOVE^{E}	5	.005	.01	-	**10**	5	.005	.01	-	**1**
Component-Cue^{E}	-	-	.005	-	**10**	-	-	.005	-	**1**
Parameter b										
ALCOVE^{L}	5	.005	.01	**3**	-	5	.005	.01	**1**	-
Component-Cue^{L}	-	-	.005	**3**	-	-	-	.005	**1**	-

Table 5.1 – *List of parameters used to produce Figure 5.2a and 5.2b. Both a high and a low values of each parameter were selected. The changing parameters are highlighted in bold letters.*

Results

Figure 5.2a and 5.2b show how a high or low parameter influence the learning curves. The magnitude of the parameter (i.e., high or low) is displayed on the columns, while the models including the parameter are displayed on the rows. The graphs highlight the category membership of the stimuli: in dark blue stimuli belonging to category A and in light blue stimuli belonging to category B. The x-axis represents the number of learning blocks, while the y-axis represents the probability of classifying the current stimulus into category A. Therefore, learning occurs when the category A probability of dark blue stimuli approaches 1 and the one of light blue stimuli approaches 0. Let us analyze the way each parameter affects the shape of the learning curves.

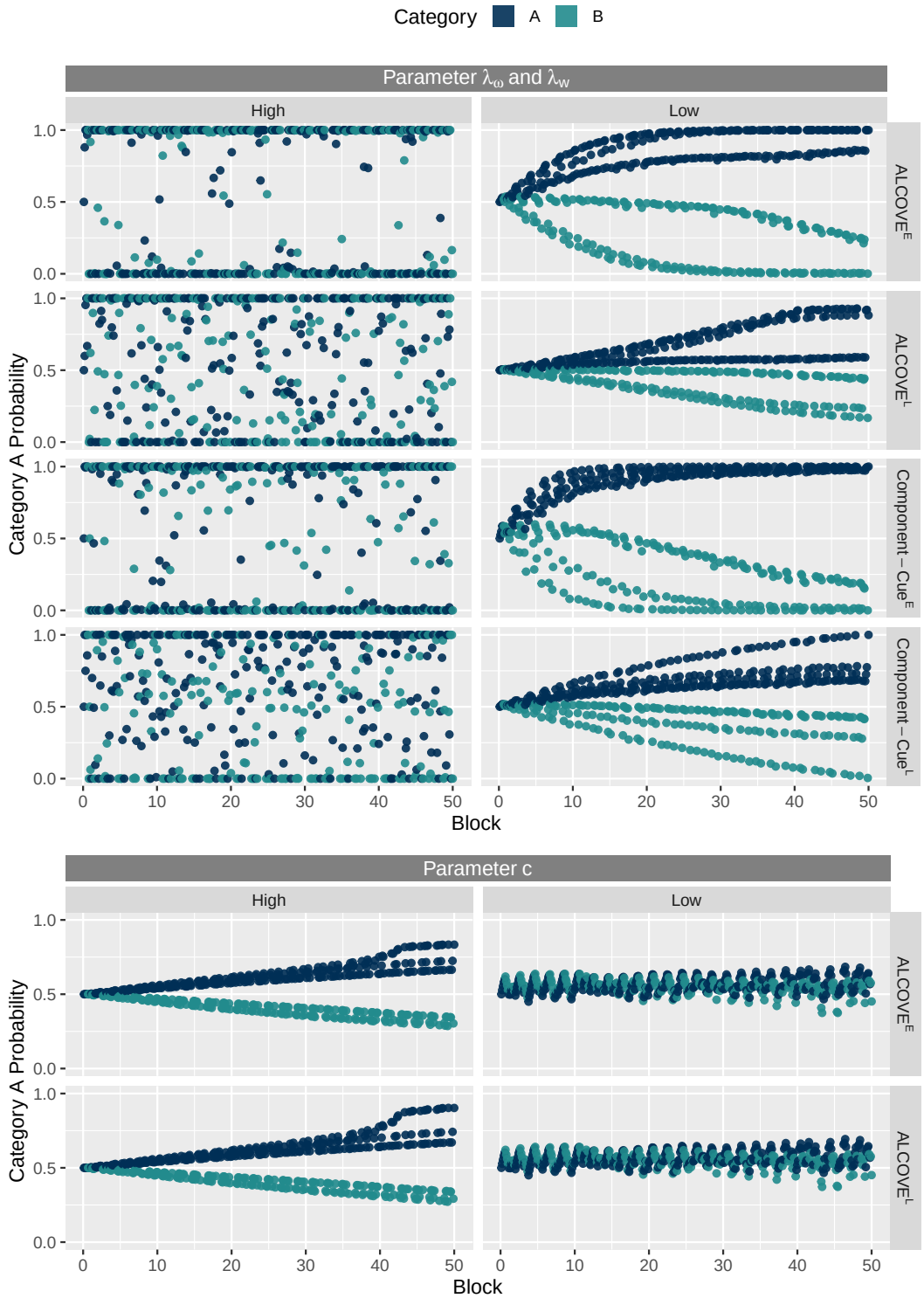


Figure 5.2a – *Variations of learning curves depending on the value of the parameters. A high and low values of each parameter is selected for the models including the parameter (the other parameters were fixed). The list of parameters used are shown in Table 5.1.*

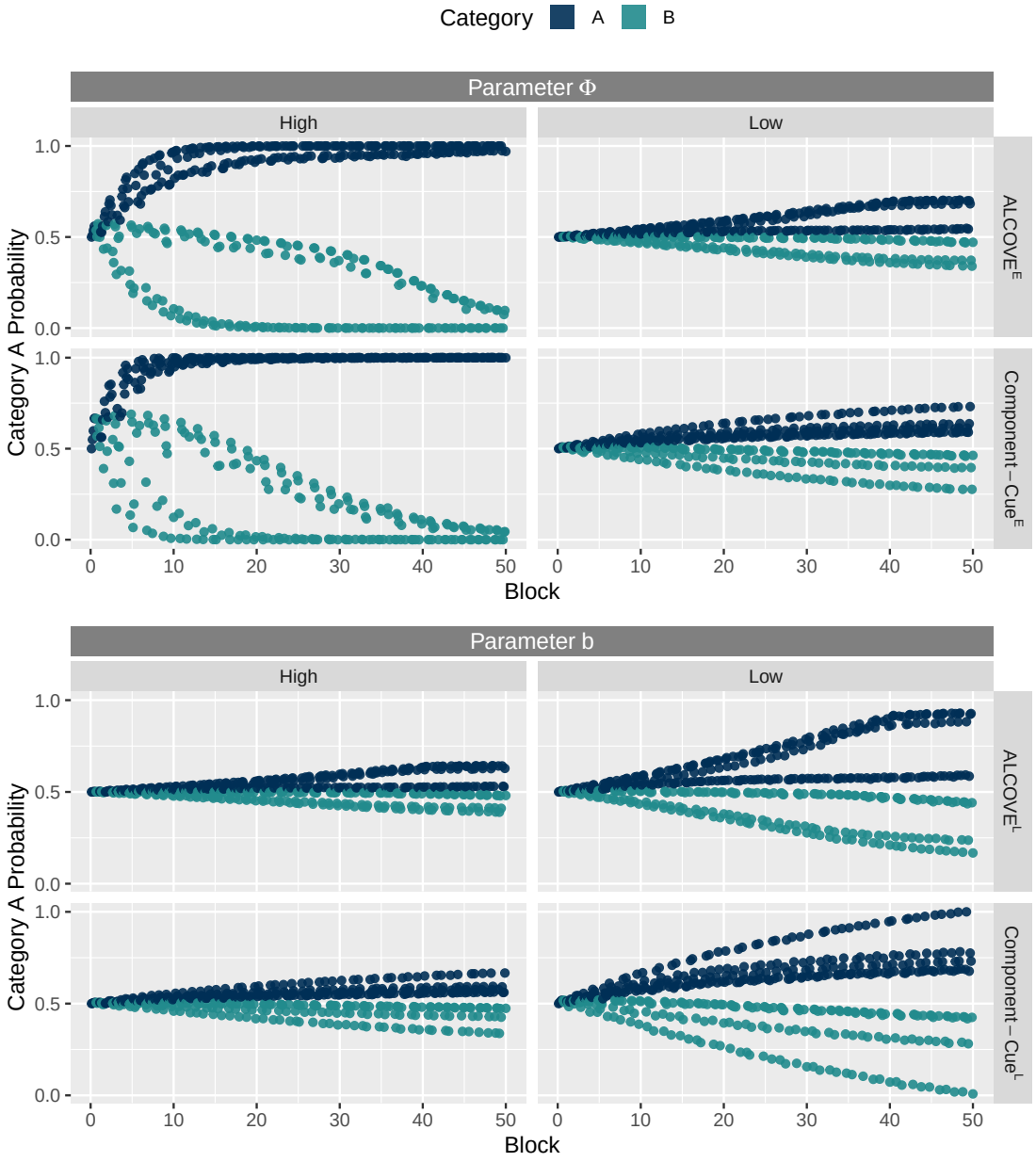


Figure 5.2b – *Variations of learning curves depending on the value of the parameters. A high and low values of each parameter is selected for the models including the parameter (the other parameters were fixed). The list of parameters used are shown in Table 5.1.*

Parameters λ_ω and λ_w. The value of the parameters λ_ω and λ_w influenced the variability of the learning curves (we recall that these parameters correspond to the magnitude of the steps during the gradient descent algorithm). High values produced highly variable learning curves, while low values produced well-defined learning curves.

Parameter c. The value of the parameter c seemed to alter both the variability of the learning curves and the value of the classification probability. More specifically, high values of c produced learning curves with small variability and high probability to correctly classify the stimuli. Conversely, low values of c produced learning curves with high variability and smaller probability to correctly classify the stimuli (as compared to the high values). This is plausible since the parameter c regulates the perception of the stimuli similarity.

Parameter ϕ. The value of the parameter ϕ affected the values of the classification probability. High values amplified the value of the classification probability. Therefore, increasing the value of ϕ increased or reduced the classification probability, depending whether the classification probability was greater or smaller than 0.5.

Parameter b. The value of the parameter b allowed the model to shrink the classification probability to a limited area centered around 0.5. The higher the value of b, the higher the compression of the learning curves to a limited area.

5.2 Parameter Estimation

As seen in the previous chapter (see Section 4.3), the consistency of the maximum likelihood estimation is not guaranteed when models are not identifiable. Because of the complex architecture of learning models, it is unclear whether it is possible to recover the values of their parameters from an infinite number of observations (i.e., whether the models are identifiable). Moreover, the fact that the predictions of learning models depend on the stimuli and feedback received before the current time warrants even more the need to evaluate the consistency of the MLE.

5.2.1 Simulated Learning Data Analysis

In this subsection, the accuracy of the maximum likelihood estimator as a function of the size of the dataset is evaluated. The procedure used on transfer models (Procedure Summary 4.3) is adapted to learning models and applied to them. Let us describe the variations in the procedure. In Step #2, the probability of classifying a stimulus into category A is also dependent on the past. Therefore, instead of $\mathbb{P}_M^\theta\left(A \,|\, x^{(i)}\right)$, we have:

$$\mathbb{P}_M^\theta\left(A \,|\, x^{(i)}, \mathcal{H}_{i-1}\right),$$

where $\mathcal{H}_{i-1}$ denotes the sequences of stimuli and feedback until time $i-1$. Similarly, in Step #4 instead of $\mathbb{P}_M^\theta\left(A \,|\, \xi_i\right)$, we have

$$\mathbb{P}_M^\theta\left(A \,|\, \xi_i, \mathcal{H}_n\right),$$

with $\mathcal{H}_n$ denoting the sequences of stimuli and feedback received during the whole learning phase (n represents the number of stimuli of the learning phase).

Technical Aspects

i. The analysis included the following models: Component-Cue$^{\mathrm{L}}$, Component-Cue$^{\mathrm{E}}$, ALCOVE$^{\mathrm{L}}$ and ALCOVE$^{\mathrm{E}}$.

ii. The studied categories were the 5-4 category set of Experiment I (see Figure 2.1). Since the investigation involved the learning phase, only the 9 learning items were considered.

iii. The accuracy of the estimation was tested on datasets characterized by the following lengths: 10, 40, 80, 120 and 160 (random) blocks of 9 items. For each length, the procedure was iterated 100 times (the generator parameter was different each time). The generator parameter was randomly selected among those generating well-defined learning curves.

iv. The gradient descent algorithm in the MLE was performed 10 times, each time starting from a random starting point.

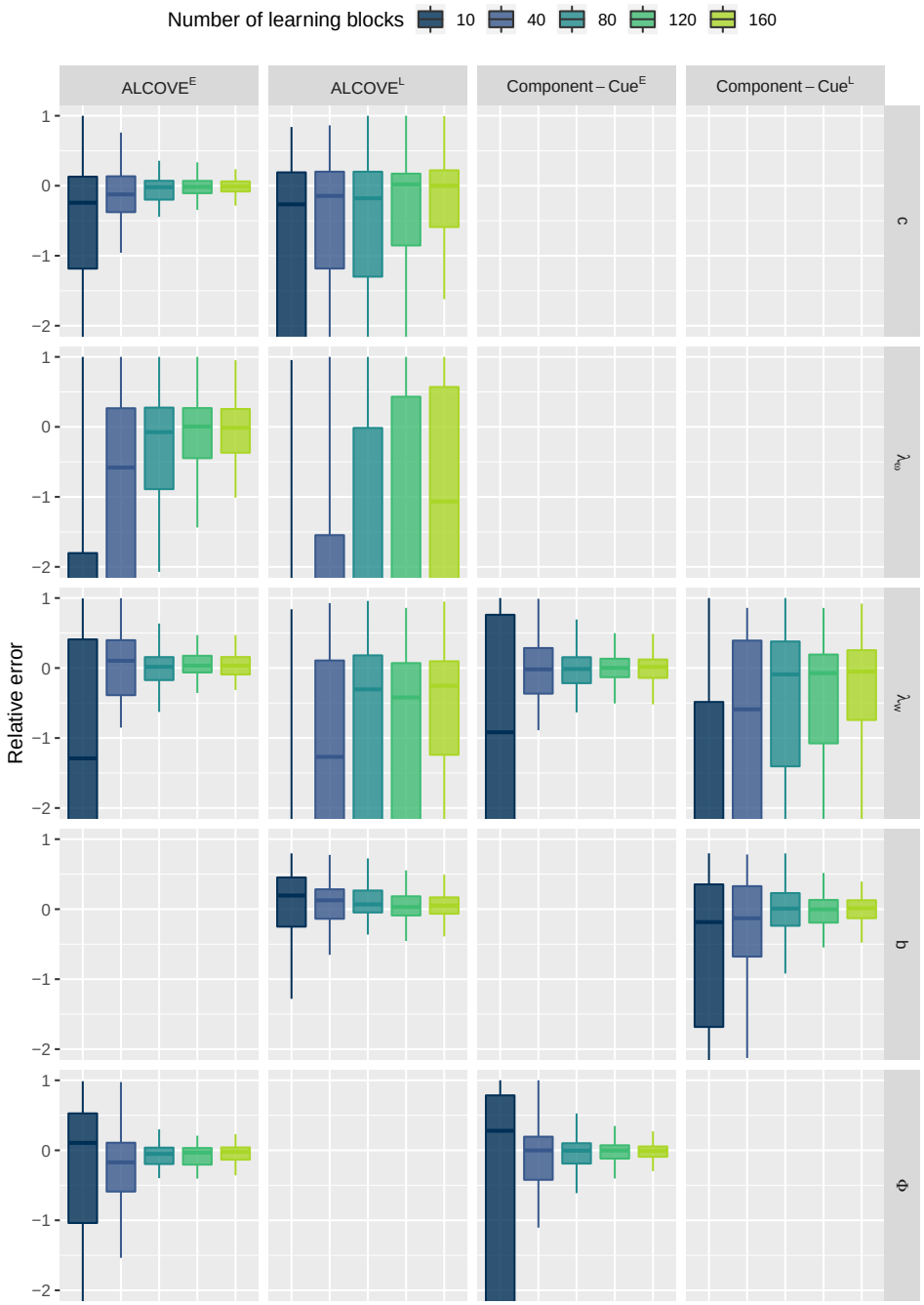


Figure 5.3 – *Box-plots representing the relative error of the maximum likelihood estimation on simulated learning data, as a function of the size of the data-set, the studied learning model, and its parameters. The relative error is defined as the difference between the generator parameter and the estimated one divided by the absolute value of the generator parameter. The box-plots were computed across 100 iterations and the generator parameter was randomly chosen at each iteration among those generating well-defined learning curves. The same items and categories of Experiment I were considered. The gradient descent algorithm in the MLE was performed 10 times.*

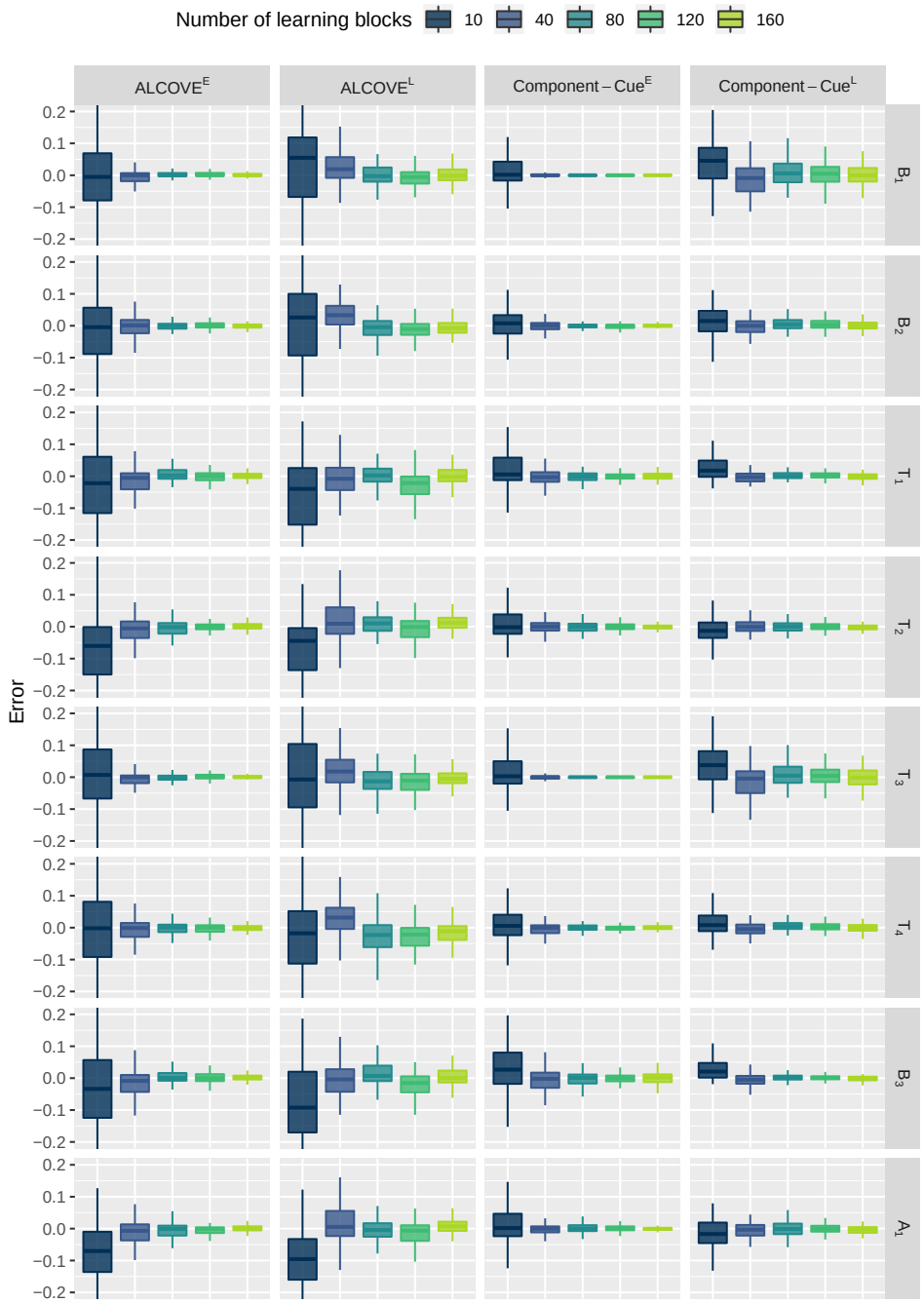


Figure 5.4a – *Box-plots representing the error between the final probability pattern computed with the generator parameter and the one computed with the estimated one (on simulated learning data), as a function of the size of the data-set, the learning model, and the items. The box-plots were computed across 100 iterations and the generator parameter was randomly chosen at each iteration among those generating well-defined learning curves. The same items and categories of Experiment I were considered. The plot involves the first 8 items of Experiment I (the last 8 are shown in Figure 5.4b). Items are denoted using the same notation as in Figure 2.1. The gradient descent algorithm in the MLE was performed 10 times.*

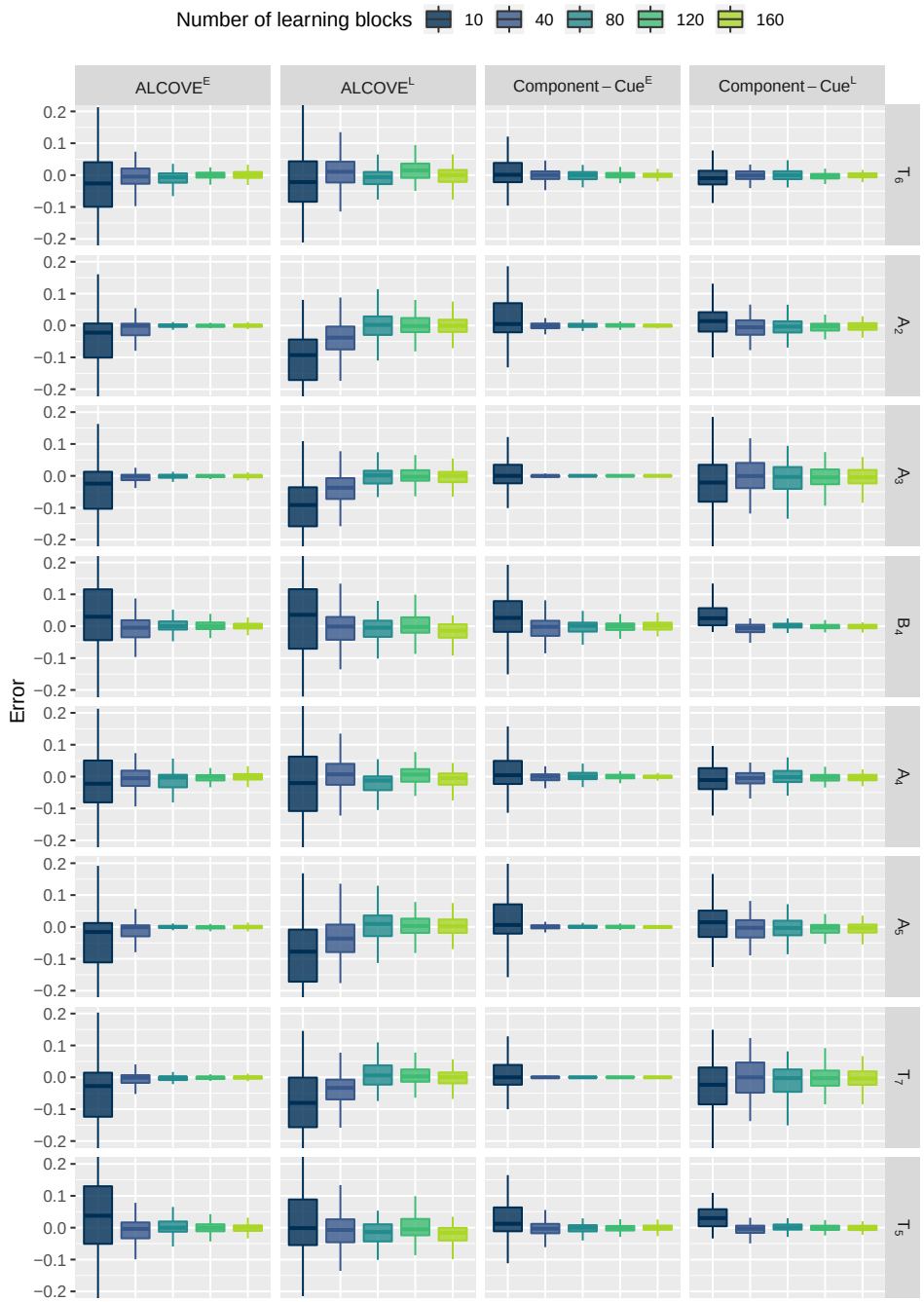


Figure 5.4b – *Box-plots representing the error between the final probability pattern computed with the generator parameter and the one computed with the estimated one (on simulated learning data), as a function of the size of the data-set, the learning model, and the items. The box-plots were computed across 100 iterations and the generator parameter was randomly chosen at each iteration among those generating well-defined learning curves. The same items and categories of Experiment I were considered. The plot involves the last 8 items of Experiment I (the first 8 are shown in Figure 5.4a). Items are denoted using the same notation as in Figure 2.1. The gradient descent algorithm in the MLE was performed 10 times.*

Results

Figure 5.3 shows the relative error between he generator parameter and the estimated one, as a function of the size of the learning data-set, the learning model, and its parameters. The learning models are displayed on the columns, while their parameters are displayed on the rows. As the number of blocks increased, the variability of the relative error decreased and the average relative error was closer to 0. The accuracy of the estimation seemed to be highly dependent on whether the exponential or linear version was considered.

Exponential versions. With the exception of the parameter λ_ω, the parameter estimation seemed to be accurate when the size of the dataset was equal to or greater than 80 blocks.

Linear versions. The accuracy of the estimation on the linear versions was not as good as the one on the exponential versions. The estimation of the parameter b seemed to be accurate when the size of the dataset was equal to or greater than 40 blocks, while the estimation of the other parameters (with the exception of λ_ω) needed 160 blocks to be accurate.

Figure 5.4a and 5.4b show the error between the ending probability pattern computed with the generator parameter and the one computed with the estimated parameter, as a function of the size of the learning data-set, the learning model, and the item. The learning models are displayed on the columns, while the items are displayed on the rows. Items are denoted using the same notation as in Figure 2.1. Again, as the number of blocks increased, the variability of the error decreased and the average error was closer to 0. Moreover, the probabilities at the end of the learning phase were accurately estimated, although some inaccuracies in the parameter estimation. The probability estimation seemed to be accurate when the size of the dataset was equal to or greater than 40 blocks.

5.3 Model Selection

In the previous chapter, we mentioned in multiple occasions (see Section 4.2 and 4.4) why cross-validation methods were preferred to probabilistic statistical criteria such as AIC or BIC. Indeed, cross-validation methods prevent the risk of overfitting the data

without penalizing the models on the basis of their number of parameters. The hold-out method is the cross-validation method that was selected to compare learning models. Since the predictions of the learning models are dependent on the entire learning process, the hold-out method is the only cross-validation method adapted to this condition. Let us describe why more robust and sophisticated cross-validation methods are inaccessible.

The learning models predict the category membership of stimuli on the basis of previous feedback. The fact that the predictions are not independent makes it difficult to apply the MLE to non-contiguous observations. Let us give an example to describe the obstacles that emerge from the application of the MLE on non-contiguous (and dependent) observations. Let us suppose that only one every two observations is accessible. With a transfer model, the MLE is easily computed because of the independence of the predictions. However, with a learning model, the computation of the MLE requires to take into account all possible values of all unseen observations. Given n unseen observations and p possible values for each observation, p^n values have to be computed. Therefore, the computational cost is the main obstacle to the implementation of the MLE on learning models. Although there are multiple techniques that allow the reduction of the computational cost, the hold-out method offered the best compromise in terms of computational cost and robustness.

5.3.1 Hold-Out Method

The hold-out method was described in Subsection 4.4.1. However, the method is adapted here to learning models. In Step #3, the probability of classifying a stimulus into category A is dependent on the past. Therefore, instead of $\mathbb{P}_M^{\theta_M}\left(A \mid x^{(i)}\right)$, we have:

$$\mathbb{P}_M^{\theta_M}\left(A \mid x^{(i)}, \mathcal{H}_{i-1}\right),$$

where $\mathcal{H}_{i-1}$ denotes the sequences of stimuli and feedback until time $i-1$. Similarly, the probability is dependent on the past also in Step #4 (Equation 4.2).

5.3.2 Validation of the Hold-Out Method

In the previous chapter, we emphasized on the importance of establishing whether the models are identifiable via the selected cross-validation method. The procedure that was applied to learning models to evaluate whether models are identifiable via the hold-out

method is similar to the procedure applied to transfer models (Subsection 4.4.3). Let us describe the few differences between the two procedures. Firstly, in Step #1, the probability of classifying a stimulus into category A is dependent on the past. Secondly, in Step #2, the hold-out method is applied instead of the k-fold cross-validation technique.

5.3.3 Simulated Data Analysis

The aim of this subsection is to determine whether learning models are identifiable via the hold-out method using the previously described procedure. The procedure was applied to two simulated datasets: one with the same features as Experiment I and the other with the same features as Experiment II. To clarify our technical choices, we need to anticipate how the hold-out method is applied to Experiment I and II. In both experiments, the hold-out method is applied to each participant. In Experiment I, the learning phase is used as the training set, while the transfer phase is used as the testing set. In Experiment II, the early 80% of the learning phase is used as the training set and the remaining 20% as the testing set.

Experiment I

Technical Aspects

i. The set of models $\mathfrak{M}$ included the following models: Component-Cue$^{\mathrm{L}}$, Component-Cue$^{\mathrm{E}}$, ALCOVE$^{\mathrm{L}}$ and ALCOVE$^{\mathrm{E}}$.

ii. The same sequence of stimuli as in Experiment I was used.

iii. The hold-out method was separately applied to each participant. The learning phase (of each participant) was used to estimate the parameters of the models, while the transfer phase (of each participant) was used to evaluate the models.

iv. The parameters used to simulate the responses are the parameters resulting from the application of the MLE to the learning phase of Experiment I.

v. The procedure was only iterated 20 times (to limit the computational cost). Thus, for each model, 20 sequences of responses associated to the sequence of stimuli of Experiment I were generated. Since Experiment I counts 43 participants, a total of 20 iterations $\times$ 43 participants $= 860$ hold-out methods were run.

vi. At each MLE, the gradient descent algorithm was performed 10 times, each time starting from a random starting point.

Results

The result of the application of the procedure to learning models on simulated data having the same features as Experiment I is shown in Figure 5.5 and 5.6. The first graph shows the number (and percentage) of times that the considered models obtained the lowest evaluation (by using the SSD or the likelihood), giving a specific simulated data sample. The second graph adopts a different prospective and shows the percentage of times that the simulated data were actually generated by the model that has the lowest evaluation (by means of the SSD or the likelihood). Although the two graphs give a complementary insight on the identifiability of the models via the hold-out method, the second graph is more useful to interpret the results on experimental data.

Let us analyze Figure 5.5. For the ALCOVE models and the exponential version of the Component-Cue model, the choice of the criterion (i.e., SSD or likelihood) did not remarkably affect the result. Conversely, for the linear version of the Component-Cue model, the choice of the likelihood criterion increased the identifiability of the model. The graph showed some important information. Firstly, Component-Cue data (i.e., both Component-Cue^E and Component-Cue^L) were more likely to be correctly identified as compared to ALCOVE data (i.e., both ALCOVE^E and ALCOVE^L). Secondly, ALCOVE data were more often misidentified as coming from the other ALCOVE version (exponential or linear) than as coming from the Component-Cue models. Moreover, when ALCOVE data were misidentified as Component-Cue data, it was more likely that the version (exponential or linear) of the generative model was the same as the selected model. The same was true for data generated with the linear version of the Component-Cue model. Conversely, the exponential Component-Cue data were more often misidentified as coming from the exponential ALCOVE than as coming from the linear Component-Cue (the linear Component-Cue is highly different from the other models).

Figure 5.6 confirmed that the Component-Cue$^{\mathrm{L}}$ was the most recognizable model with almost 90% of chance to be correctly identified. The ALCOVE models had more than 80% of chance to be recognized as such, but they were identified with the incorrect version 25% of time approximately. Conversely, the Component-Cue$^{\mathrm{E}}$ had higher probability to be correctly recognized as compared to the ALCOVE models (approximately 64%-65%

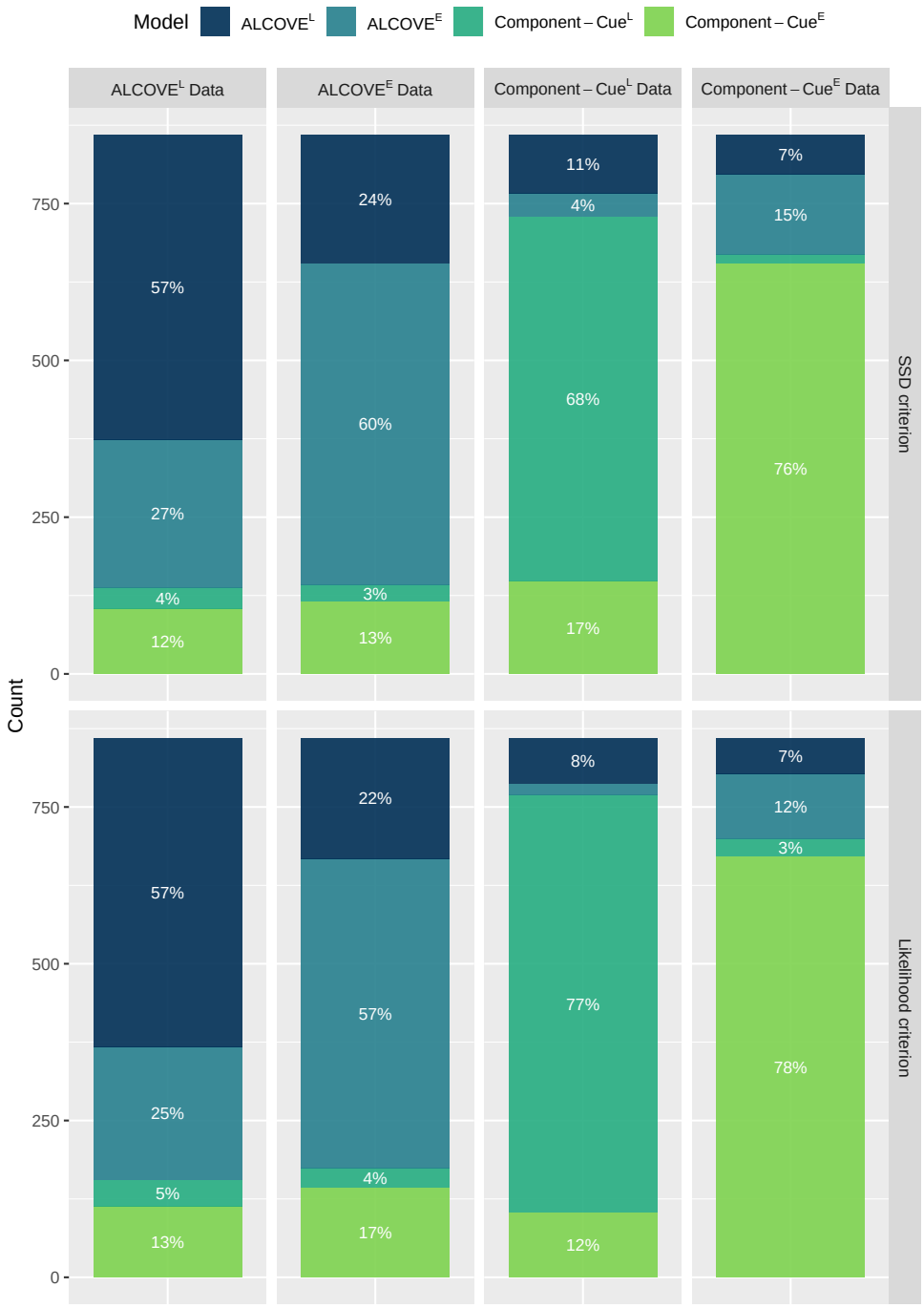


Figure 5.5 – *Validation of the hold-out method on simulated data. The graph shows the number of times the samples generated by specific models are actually recognized as such. The same sequence of stimuli of Experiment I was used. The procedure was iterated 20 times for a total of 20 iterations* $\times$ *43 participants* $= 860$ *hold-out methods. The hold-out method was applied to each participant, separately. The parameter estimation was performed on the learning phase, while the evaluation of the model was performed on the transfer phase. Models are evaluated with both the SSD and the likelihood. The gradient descent algorithm in the MLE was performed 10 times.*

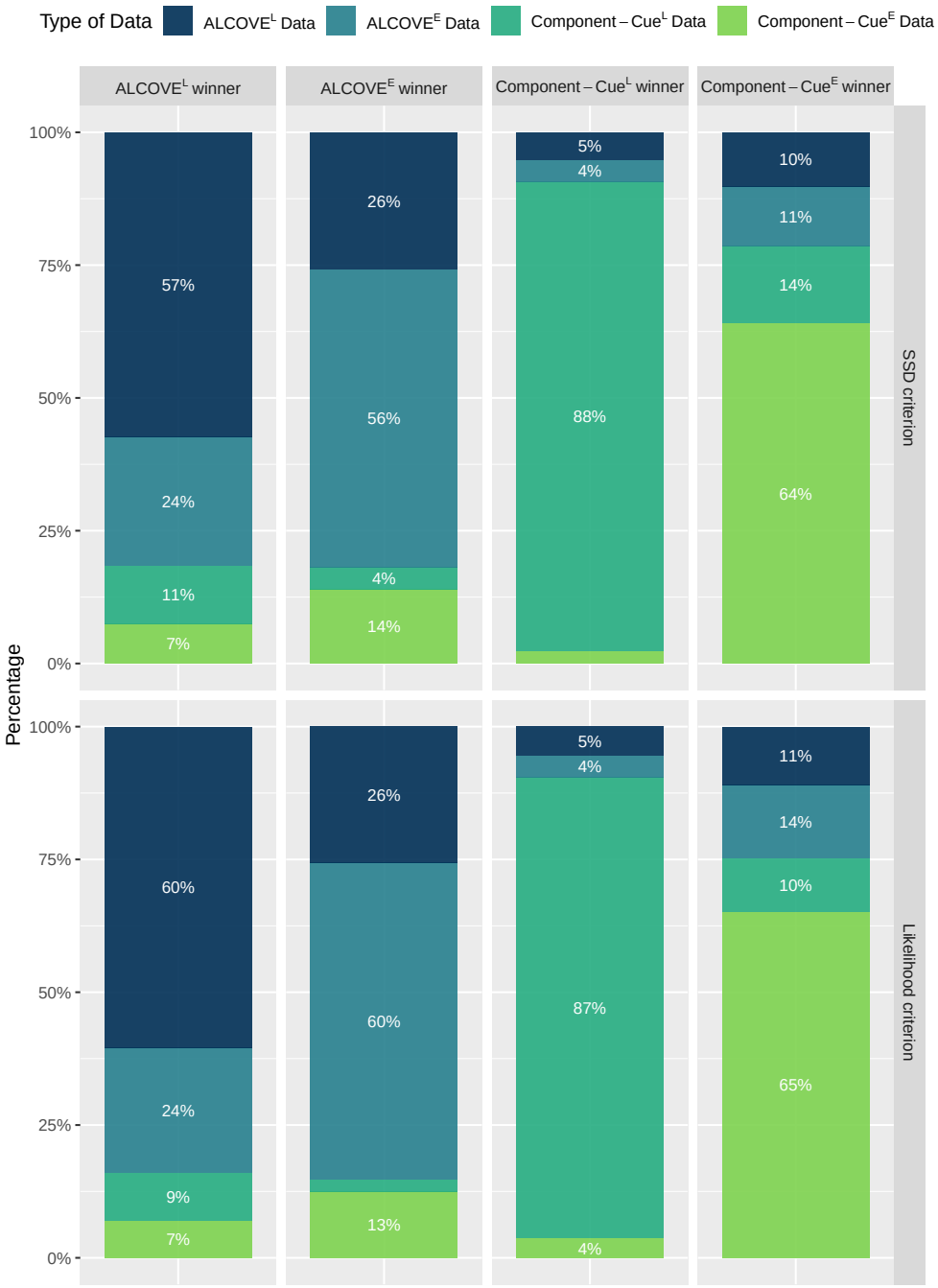


Figure 5.6 – *Validation of the hold-out method on simulated data. The graph shows the percentage of times that the model with the lowest evaluation is actually the generative model of the data. The same sequence of stimuli of Experiment I was used. The procedure was iterated 20 times for a total of 20 iterations* $\times$ *43 participants* $= 860$ *hold-out methods. The hold-out method was applied to each participant, separately. The parameter estimation was performed on the learning phase, while the evaluation of the model on the transfer phase. Models are evaluated with both the SSD and the likelihood. The gradient descent algorithm in the MLE was performed 10 times.*

instead of 56%-60%), but smaller probability to be recognized as a Component-Cue, regardless of the version (75%-78% instead of 82%-86%).

To recap, if the model that has the lowest evaluation is Component-Cue^L, then it is the generative model with a probability of 87-88%. If the model that has the lowest evaluation is Component-Cue^E, then it is the generative model with a probability of 64-65%. Finally, if the model that has the lowest evaluation is either ALCOVE^L or ALCOVE^E, then it is the generative model with a probability of 57-58%.

Experiment II

Technical Aspects

i. The set of models $\mathfrak{M}$ included the following models: Component-Cue$^{\mathrm{L}}$, Component-Cue$^{\mathrm{E}}$, ALCOVE$^{\mathrm{L}}$ and ALCOVE$^{\mathrm{E}}$.

ii. The same sequence of stimuli as in Experiment II was used.

iii. The hold-out method was separately applied to each participant. For each participant, the early 80% of the learning phase was used to estimate the parameters of the models, while the remaining 20% was used to evaluate the models.

iv. The parameters used to simulate the responses are the parameters resulting from the application of the MLE to the learning phase of Experiment II.

v. The procedure was only iterated 20 times (to limit the computational cost). Thus, for each model, 20 sequences of responses associated to the sequence of stimuli of Experiment II were generated. Since Experiment I counts 22 participants, a total of 20 iterations $\times$ 22 participants $= 440$ hold-out methods were run.

vi. At each MLE, the gradient descent algorithm was performed 10 times, each time starting from a random starting point.

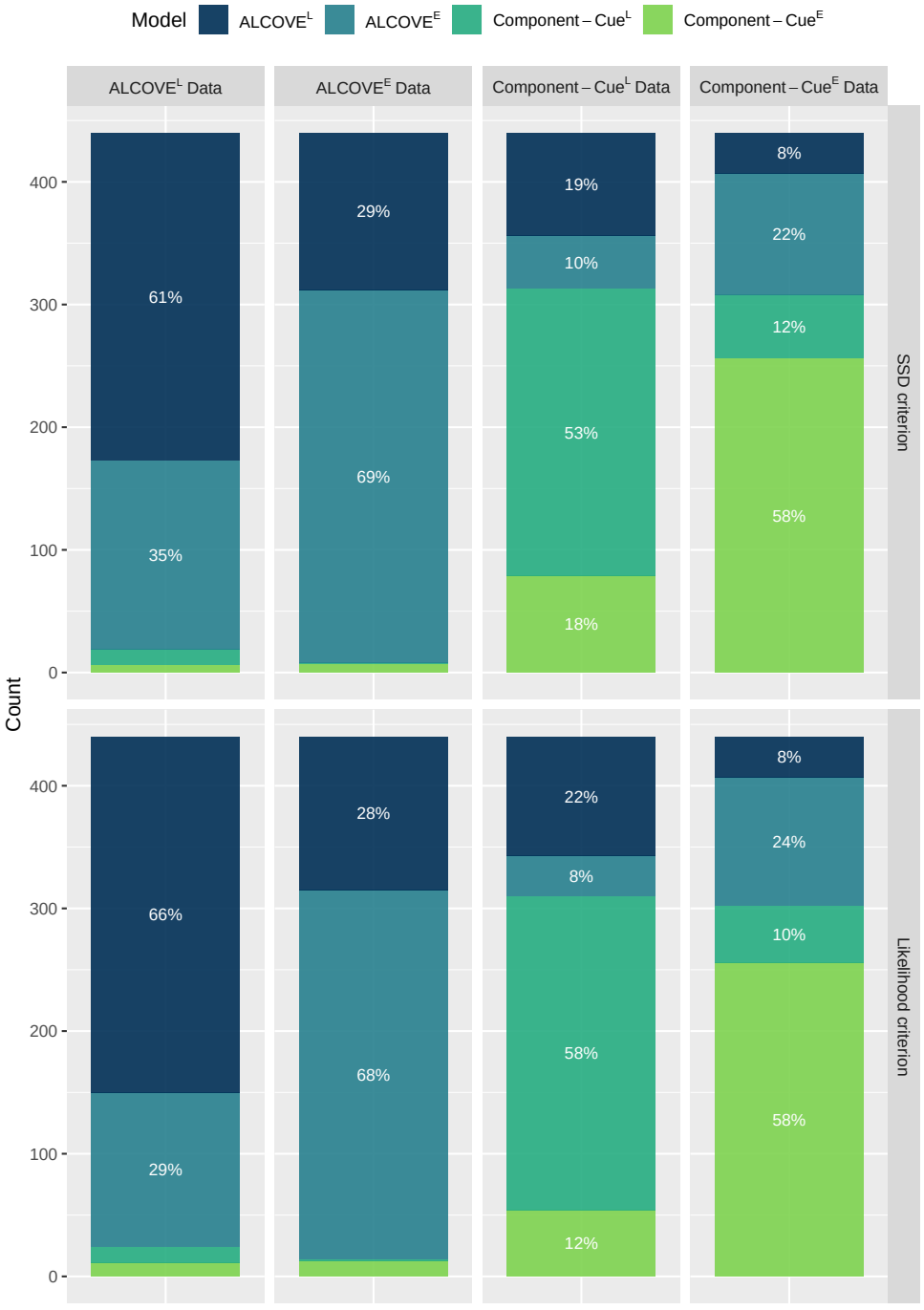


Figure 5.7 – *Validation of the hold-out method on simulated data. The graph shows the number of times the samples generated by specific models are actually recognized as such. The same sequence of stimuli of Experiment II was used. The procedure was iterated 20 times for a total of 20 iterations* $\times$ *22 participants* $= 440$ *hold-out methods. The hold-out method was applied to each participant, separately. The parameter estimation was performed on the early 80% of the learning phase, while the evaluation of the model was performed on the remaining 20%. Models are evaluated with both the SSD and the likelihood. The gradient descent algorithm in the MLE was performed 10 times.*

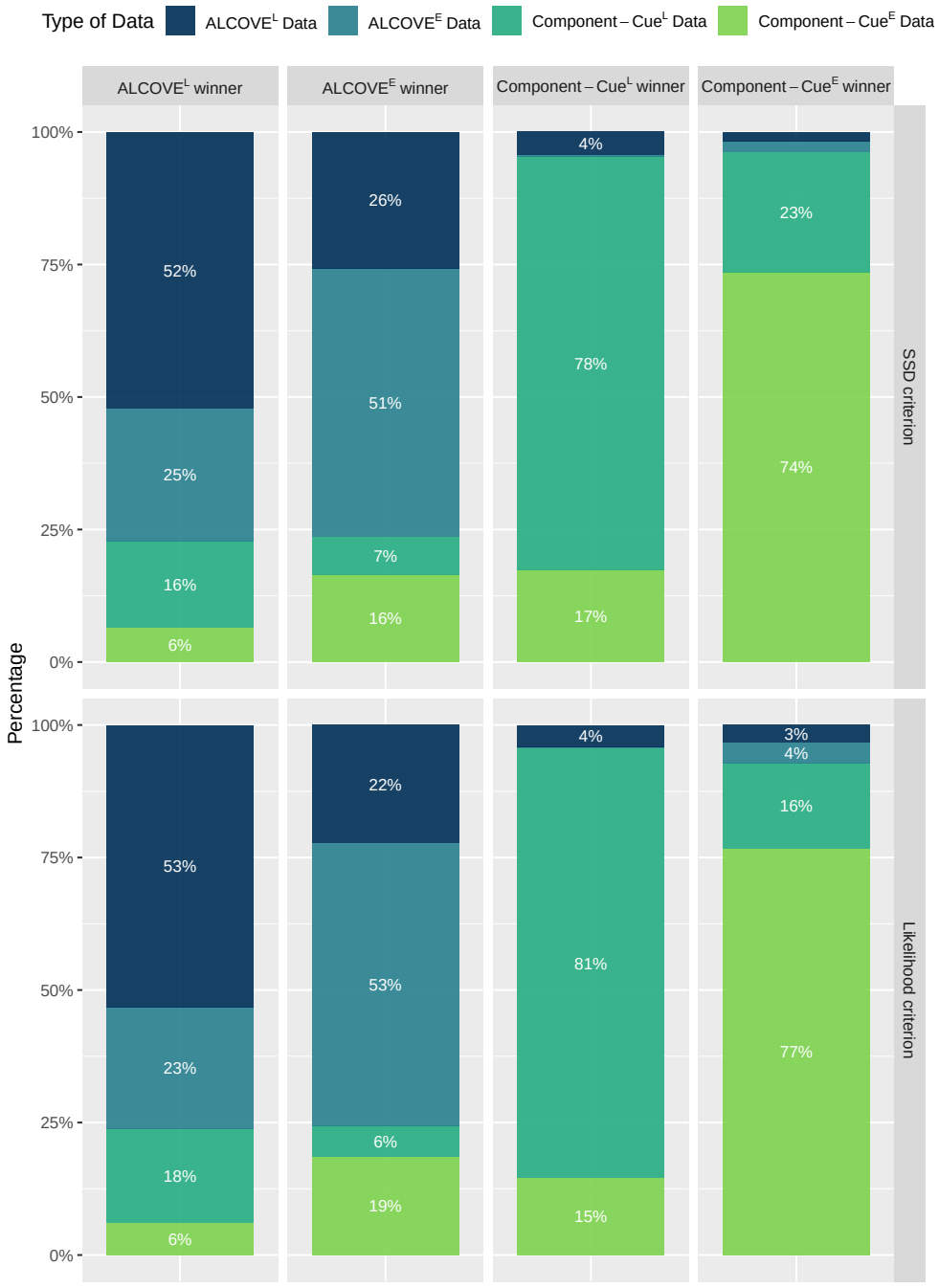


Figure 5.8 – *Validation of the hold-out method on simulated data. The graph shows the percentage of times that the model with the lowest evaluation is actually the generative model of the data. The same sequence of stimuli of Experiment II was used. The procedure was iterated 20 times for a total of 20 iterations* $\times$ *22 participants* $= 440$ *hold-out methods. The hold-out method was applied to each participant, separately. The parameter estimation was performed on the early 80% of the learning phase, while the evaluation of the model on the remaining 20%. Models are evaluated with both the SSD and the likelihood. The gradient descent algorithm in the MLE was performed 10 times.*

Results

The result of the application of the procedure to the learning models on simulated data having the same features as Experiment II are shown in Figure 5.7 and 5.8. Again, the first graph illustrates the number (and percentage) of times that samples generated with a specific model were actually recognized as such. The second graph shows the percentage of times that the model with the lowest evaluation (by means of the SSD or the likelihood) was actually the generative model of the simulated data.

Let us analyze Figure 5.7. ALCOVE data have smaller probability to be misidentified as Component-Cue data with respect to Figure 5.5, while Component-Cue data have greater probability to be misidentified as generated from ALCOVE models with respect to Figure 5.5.

Figure 5.8 shows that, when either the exponential or linear version of Component-Cue has the lowest evaluation, then the probability that Component-Cue was the generative model is 93-96%. However, the incorrect version is identified as the generative model 15-23% of the time. Conversely, when either the exponential or linear version of ALCOVE has the lowest evaluation, then the probability that ALCOVE was the generative model is 77-78% (with 22-26% of chances the incorrect version is chosen). To conclude, the model that has the lowest evaluation is the generative model with an average probability of 79% (for the Component-Cue models) and 53% (for the ALCOVE modes).

To recap, if the model that has the lowest evaluation is either Component-Cue^L or Component-Cue^E, then it is the generative model with a probability of 77-78%. If the model that has the lowest evaluation is either ALCOVE^L or ALCOVE^E, then it is the generative model with a probability of 52-53%.

5.4 Experimental Data Analysis

The aim of this section is to apply the hold-out method (Procedure Summary 4.4) to both Experiment I and II to determine the learning models that best fits the experimental data. If the limited size of the transfer data was an obstacle during the comparative analysis of transfer models, this issue is no longer relevant. The size of the learning phase allowed us to apply the hold-out method to each participant. Thus, the data-set of each participant was divided into training and testing sets. In Experiment I, the learning phase was used

as the training set, while the transfer phase was used as the testing set. In Experiment II, the early 80% of the learning phase was used as the training set and the remaining 20% as the testing set. The choice 80-20 reflects the proportion of training-testing observations used in the 5-fold (which is one of the most commonly used k-fold).

The section is organized in four subsections. The first two subsections are devoted to the application of the hold-out method to both Experiment I and II. The last two subsections are devoted to the investigation of the relation between the within-category presentation order (rule-based vs. similarity-based) and the type of model that best fits the experimental data (ALCOVE vs. Component-Cue).

5.4.1 Analysis of Experiment I

Technical Aspects

i. The comparison of the learning models was performed on both the learning and transfer phases of Experiment I and involved the following models: Component-Cue$^{\mathrm{L}}$, Component-Cue$^{\mathrm{E}}$, ALCOVE$^{\mathrm{L}}$ and ALCOVE$^{\mathrm{E}}$.

ii. The hold-out method was applied to each participant. The learning phase was used to estimate the parameters of the models, while the transfer phase was used to test them.

iii. At each MLE, the gradient descent algorithm was performed 10 times, each time starting from a different starting point. The starting points were randomly selected from the collection of parameters satisfying the following constraints: c and ϕ were between 0 and 10; b between 1 and 2; λ_ω between 0 and 0.1; λ_w between 0 and 0.1 for the ALCOVE models and between 0 and 0.01 for the Component-Cue models. These constrains ensured that the learning curves were well-defined.

Results

The result of the application of the hold-out method to Experiment I is shown in Figure 5.9 (on the top). The graph shows the number and percentage of participants that were best fit by the learning models, depending on the selected criterion. The performance of the

majority of the participants (66% approximately) was better reproduced by Component-Cue rather than ALCOVE (with a dominance of the linear version when the SSD criterion is used and a dominance of the exponential one when the likelihood criterion is used).

In Figure 5.9 (on the bottom) the result of the application of the hold-out method to Experiment I is shown as a function of the within-category order (rule-based vs. similarity-based). The number of participants in the similarity-based order whose performance were best reproduced by ALCOVE was higher than the number of participants in the rule-based order. Conversely, the number of participants in the similarity-based order whose performance were best reproduced by Component-Cue was smaller than the number of participants in the rule-based order (statistical analyses will be performed in Subsection 5.4.3).

Figure 5.10a and 5.10b shows the (transfer) predictions of the models as a function of both the items and the within-category order. On the columns, the predictions of the models of all participants, of participants in the rule-based order, and of participants in the similarity-based order are displayed. On the rows, the items are displayed (learning items in Figure 5.10a and transfer items in Figure 5.10b). The participants' transfer performance are indicated with an x-mark. Both the predictions of the models and the participants' performance were averaged. All models achieved good quantitative predictions on the learning items (the Component-Cue^L had the worse and less extreme predictions). On the transfer items, the accuracy of the predictions were variable. All models were unable to account for the participants' performance on items T_1 and T_6 (the Component-Cue^L had the best predictions). However, they achieved good predictions on the remaining transfer items. Component-Cue^L was the model that best reproduce participants' performance on the transfer items. Finally, the predictions of the models were not influenced by the presentation order (all models provided similar predictions for participants in the rule- and similarity-based order).

Figure 5.9 – *Result of the hold-out method applied to Experiment I. The graph shows the number and percentage of participants that were best fit by the learning models, as a function of the models (shades of blue) and the evaluation criterion (i.e., SSD and likelihood; as columns). The parameter estimation was performed on the learning phase, while the evaluation of the model on the transfer phase. The gradient descent algorithm in the MLE was performed 10 times. On the bottom, the result of the hold-out method as a function of the within-category order.*

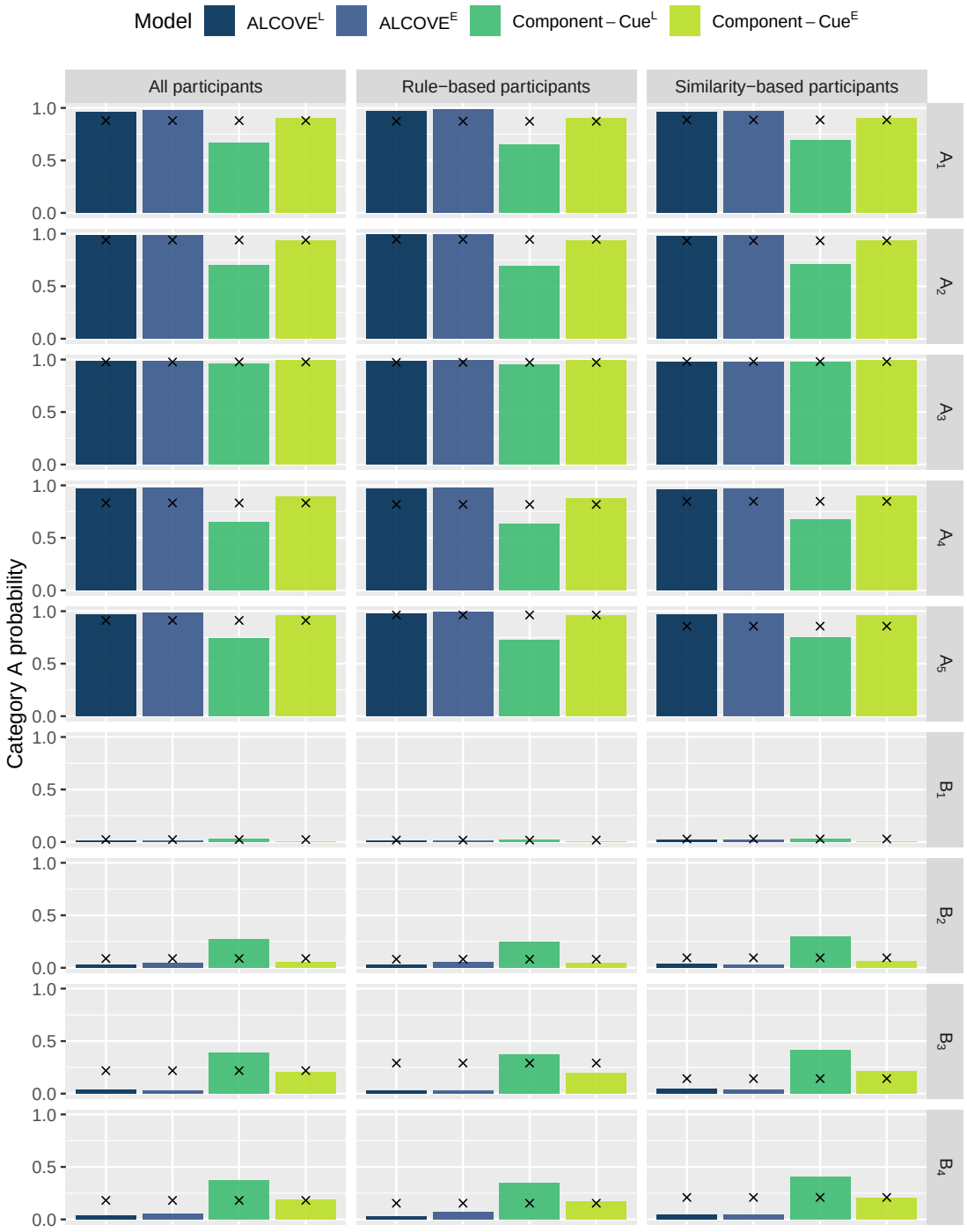


Figure 5.10a – *Predictions of the learning models on the transfer phase of Experiment I, as a function of both the items and the within-category order. Only learning items are considered. The participants' transfer performance are indicated with an x-mark. Both the predictions of the models and the participants' performance were averaged.*

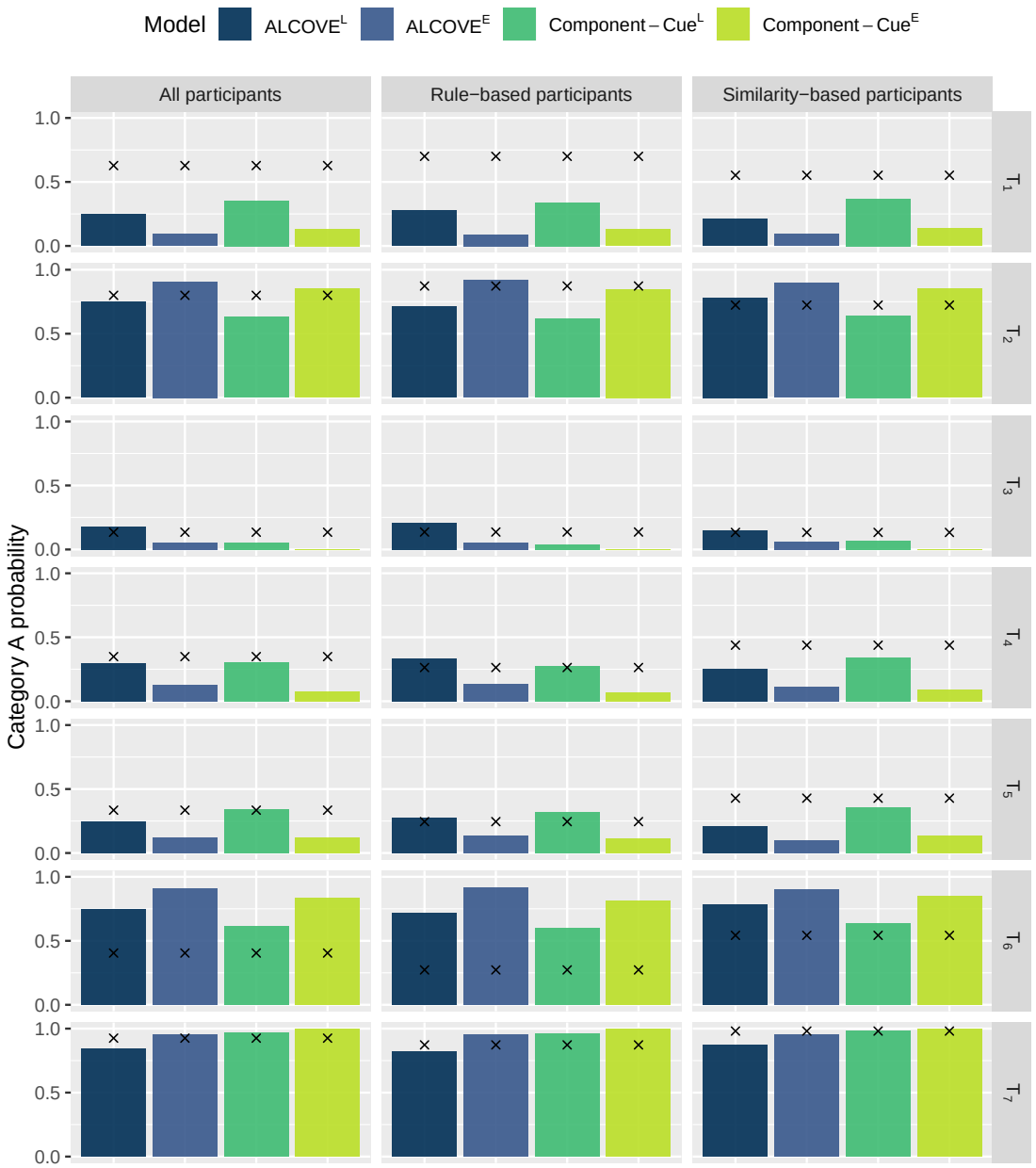


Figure 5.10b – *Predictions of the learning models on the transfer phase of Experiment I, as a function of both the items and the within-category order. Only transfer items are considered. The participants' transfer performance are indicated with an x-mark. Both the predictions of the models and the participants' performance were averaged.*

5.4.2 Analysis of Experiment II

Technical Aspects

i. The comparison of the learning models was performed on Experiment II and involved the following models: Component-Cue^{L}, Component-Cue^{E}, ALCOVE^{L} and ALCOVE^{E}.

ii. The hold-out method was applied to each participant. The early 80% of the learning phase was used to estimate the parameters of the models, while the remaining 20% was used to test them.

iii. At each MLE, the gradient descent algorithm was performed 10 times, each time starting from a different starting point. The starting points were randomly selected from the collection of parameters satisfying the following constraints: c and ϕ were between 0 and 10; b between 1 and 2; λ_ω between 0 and 0.1; λ_w between 0 and 0.1 for the ALCOVE models and between 0 and 0.01 for the Component-Cue models. These constrains ensured that the learning curves were well-defined.

Results

The result of the application of the hold-out method to Experiment II is shown in Figure 5.11. The graph shows the number and percentage of participants that were best fit by the learning models, as a function of the models (shades of blue), the evaluation criterion (i.e., SSD and likelihood; as columns), and the context (i.e., Random-Variable, Random-Constant and Blocked-Constant; as rows). Almost all participants' performance was better reproduced by ALCOVE rather than Component-Cue (with a dominance of the linear version when the SSD is used and a dominance of the exponential version when the likelihood is used).

In Figure 5.12 the result of the application of the hold-out method to the Random-Variable context is shown as a function of the within-category order (rule-based vs. similarity-based). The other contexts are not shown since almost the totality of the participants was best fit by ALCOVE. No relation between the type of model and the within-category order was visible (further investigation are conducted in Subsection 5.4.4).

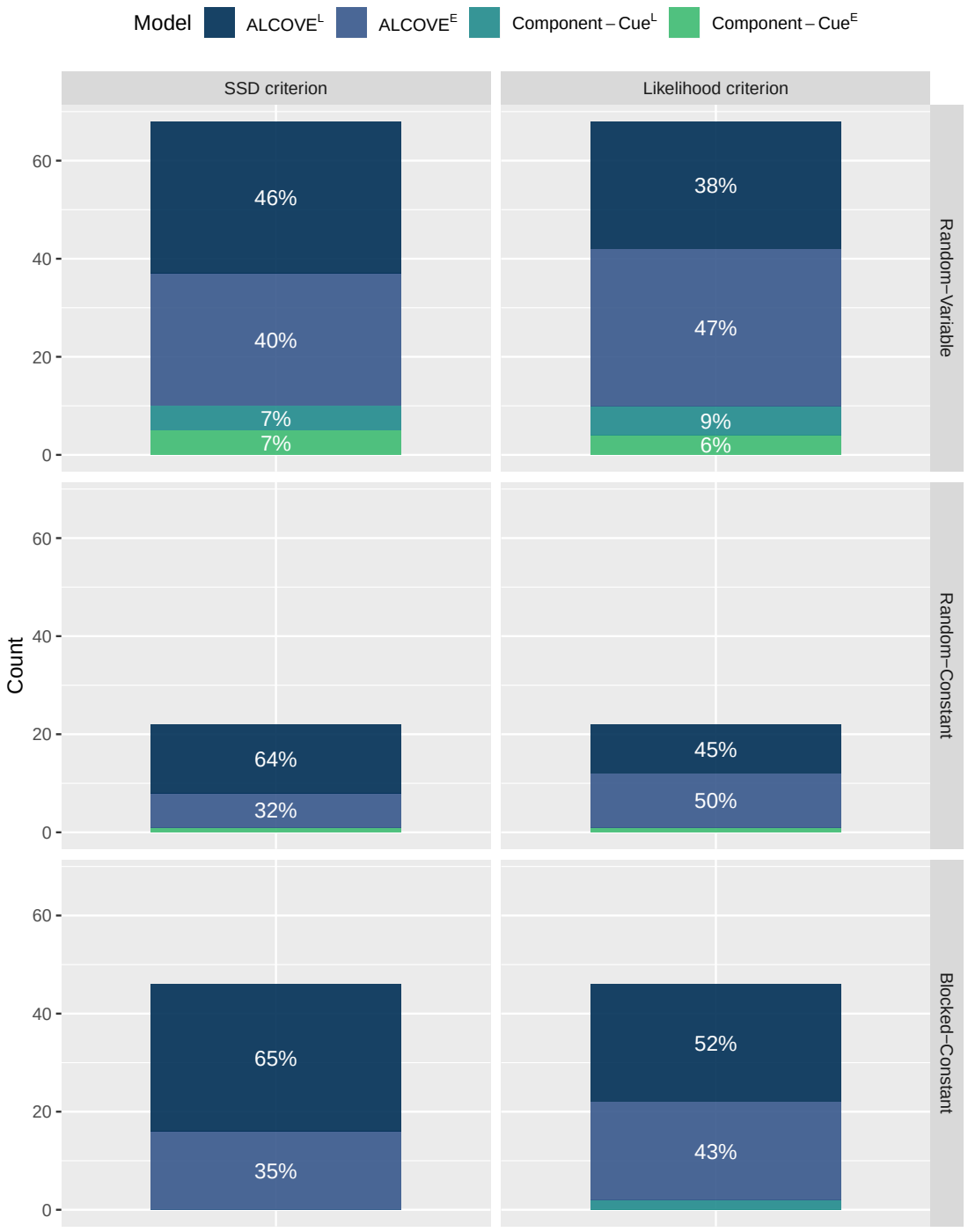


Figure 5.11 – *Result of the hold-out method applied to Experiment II. The graph shows the number and percentage of participants that were best fit by the learning models, as a function of the models (shades of blue), the evaluation criterion (i.e., SSD and likelihood; as columns), and the context (i.e., Random-Variable, Random-Constant and Blocked-Constant; as rows). The parameter estimation was performed on the early 80% of the learning phase, while the evaluation of the model on the remaining 20%. The gradient descent algorithm in the MLE was performed 10 times.*

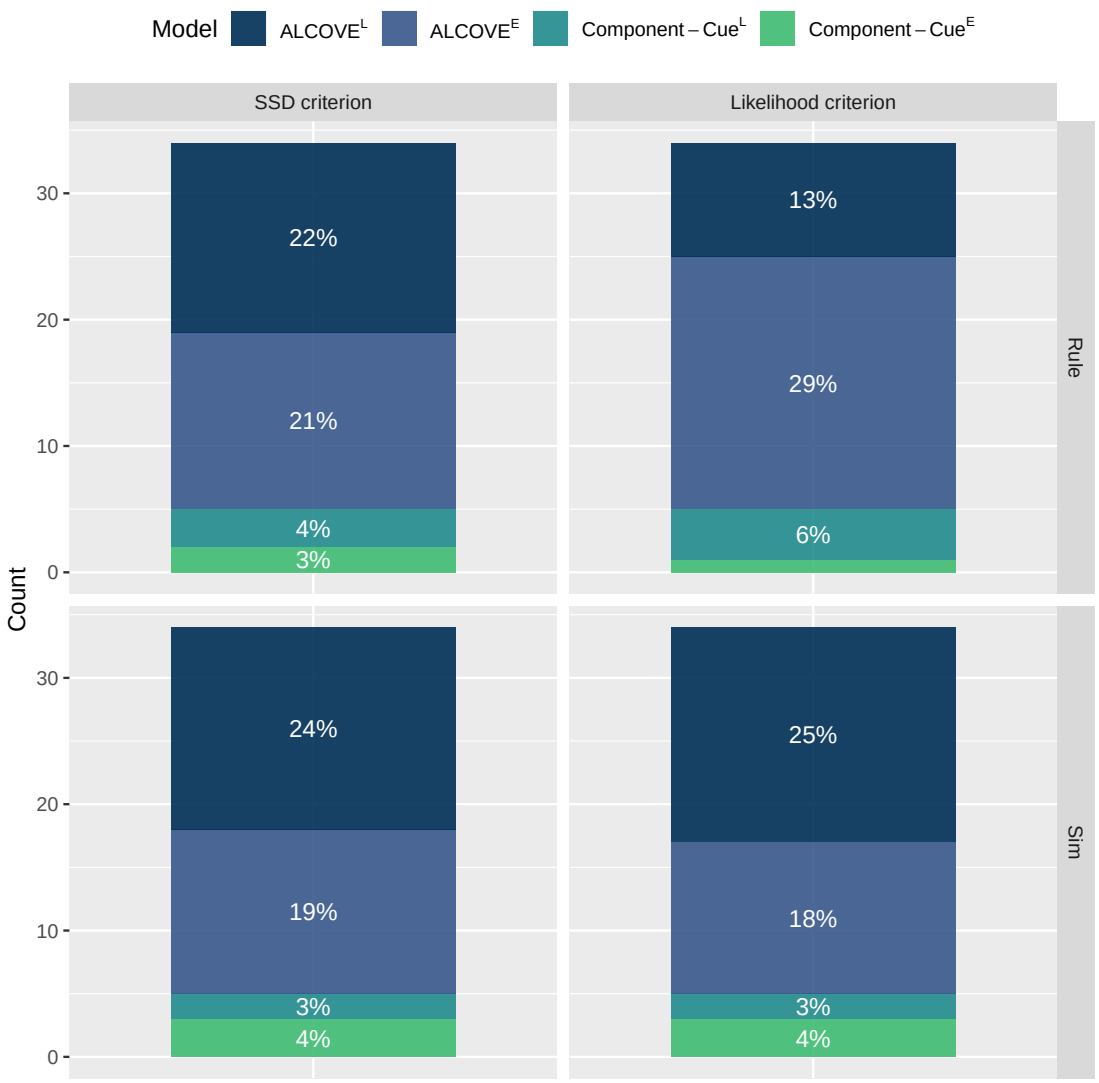


Figure 5.12 – *Result of the hold-out method applied to the Random-Variable context. The graph shows the number and percentage of participants that were best fit by the learning models, as a function of the models (shades of blue), the evaluation criterion (i.e., SSD and likelihood; as columns), and the within-category order (i.e., rule-based vs. similarity-based; as rows). The parameter estimation was performed on the early 80% of the learning phase, while the evaluation of the model on the remaining 20%. The gradient descent algorithm in the MLE was performed 10 times.*

5.4.3 Rule-Based vs. Similarity-Based from a Model Perspective in Experiment I

In Subsection 3.3.3, we mentioned that ALCOVE and Component-Cue integrate two different strategies: a similarity-based strategy for ALCOVE and a rule-based strategy for Component-Cue. Moreover, Mathy and Feldman [MF16] showed that participants in the rule-based order exhibit generalization patterns that are consistent with a rule-based strategy. In other words, the rule-based order promotes the use of a rule-based strategy. Therefore, one plausible hypothesis is that the Component-Cue model would better perform on participants in the rule-based order rather than on participants in the similarity-based order, while ALCOVE would better perform on participants in the similarity-based order rather than on participants in the rule-based order. The aim of the present subsection is to investigate this hypothesis and determine whether the within-category order (rule-based vs. similarity-based) is related to the type of model (ALCOVE vs. Component-Cue). The following aspects were investigated:

Number of participants. Firstly, we analyzed whether the within-category order (rule-based vs. similarity-based) was related to the number of participants whose responses were better predicted by a specific type of model (ALCOVE vs. Component-Cue).

Learning times. Secondly, we examined whether the time at which participants reached the learning criterion was related to the type of model that better predicted their responses.

Generalization patterns. Thirdly, we explored whether the within-category order was related to the generalization patterns of the model that better predicted the participants' responses (and its type).

Within-category order sensitivity. Finally, we investigated whether both the learning curves and the generalization patterns of the models were sensitive to the within-category order.

	SSD WINNER		LIKELIHOOD WINNER	
	ALCOVE	Component-Cue	ALCOVE	Component-Cue
Rule-based	6	16	7	15
Similarity-based	9	12	9	12

Table 5.2 – *Number of participants of Experiment I whose responses were best predicted by either the Component-Cue models or the ALCOVE models, as a function of both the within-category order (rule-based vs. similarity-based) and the evaluation criterion (SSD vs. Likelihood).*

Number of Participants

The aim is to determine whether participants in the rule-based order were best fit by Component-Cue and participants in the similarity-based order were best fit by ALCOVE.

Fisher's exact test of independence. Table 5.2 shows the number of participants whose responses were best predicted by either the Component-Cue models or the ALCOVE models, as a function of the within-category order (rule-based vs. similarity-based) and the evaluation criterion (SSD vs. Likelihood). Component-Cue best fits a higher number of participants in the rule-based order as compared to participants in the similarity-based order. Conversely, ALCOVE best fits a higher number of participants in the similarity-based order as compared to participants in the rule-based order. A Fisher's exact test of independence was performed on the SSD and likelihood tables, separately (see Table 5.2). The tests were not significant (p-value=0.35 for the SSD and p-value=0.54 for the likelihood).

Conclusion: No significant relation was found between the within-category order and the type of models.

Learning Times

In Subsection 2.1.2, we found that participants in the rule-based order had higher probability to meet the learning criterion as compared to participants in the similarity-based order. Here, we search for an (indirect) relation between within-category order and types of models by analyzing the time at which participants reached the learning

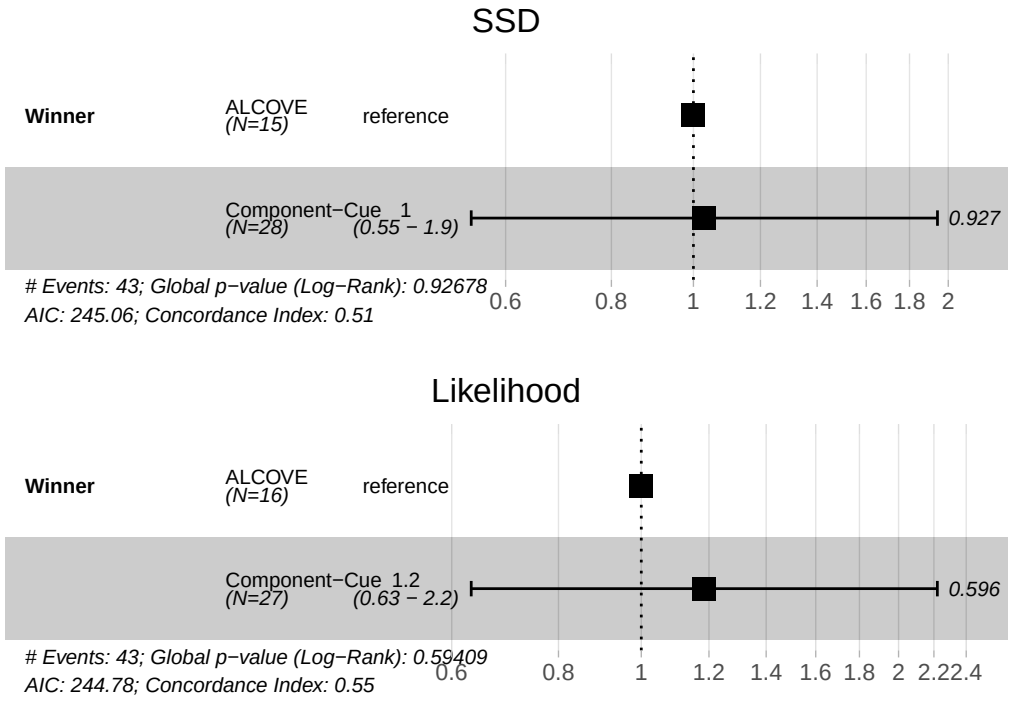


Figure 5.13 – *Result of the Cox model examining the influence of the type of winning model on the rate at which participants reached the learning criterion, for both evaluation criteria (SSD on the top and likelihood on the bottom). The technique was applied to the winning models resulting from the application of the hold-out method to Experiment I. ALCOVE is the reference condition with a hazard ratio of 1, while Component-Cue is the opposite condition. The numbers within the brackets represent the 95% confidence interval. The number on the right side of the graph is the p-value of the Wald test assessing the significance of the model.*

criterion. The aim is to determine whether participants whose responses were best predicted by Component-Cue have high probability to reach the learning criterion as compared to participants whose responses were best predicted by ALCOVE. The analysis was conducted by means of the Cox proportional-hazards model.

Cox proportional-hazards model. The results of the application of the Cox model to the SSD and likelihood criteria (separately) are displayed in Figure 5.13 (SSD on the top and likelihood on the bottom). ALCOVE is the reference condition with a hazard ratio of 1. Component-Cue has an identical hazard ratio when the SSD is considered, and a slightly higher hazard ratio when the likelihood is considered. However, in both cases

the Cox model was not significant (p-value=0.927 for the SSD and p-value=0.596 for the likelihood), showing no relation between the type of winning models and the hazard rate.

Conclusion: No relation was found between the type of winning model and the hazard rate. Therefore, no indirect relation was detected between the type of winning model and the within-category order.

Generalization Patterns

In Subsection 2.1.3, we performed a test investigating the influence of within-category order on generalization patterns (i.e., the predictions of the models after a period of training). The test showed that generalization patterns were influenced by the within-category order and that participants in the rule-based order exhibited patterns consistent with a rule-based retrieval. Here, we conduct a similar analysis examining the transfer predictions of the models that best fit participants' performance.

Principal component analysis with Wilcoxon-Mann-Whitney test. Firstly, we computed the generalization patterns of the models that best reproduced participants' performance (only transfer items are considered). Secondly, we projected these generalization patterns on the same plan as in Figure 2.9 (the patterns were also scaled in the same way before the projection).

The result of these two steps are shown in Figure 5.14, where the projection of the generalization patterns are displayed as a function of the winning model and the within-category order. On the first component, participants in the rule-based order are mostly located in the right side of the graph, while participants in the similarity-based order are mostly located in the left side of the graph (as in Figure 2.9). The one-sided Wilcoxon-Mann-Whitney performed on rule- and similarity-based participants detected a significant difference in location (p-value=0.04), showing that the winning models were able to capture the difference in generalization patterns between rule-based and similarity-based participants.

Moreover, almost all Component-Cue winners were located on the right side of the graph, while almost all ALCOVE winners were located on the left side of the graph. Again, the one-sided Wilcoxon-Mann-Whitney test confirmed that the difference in location was

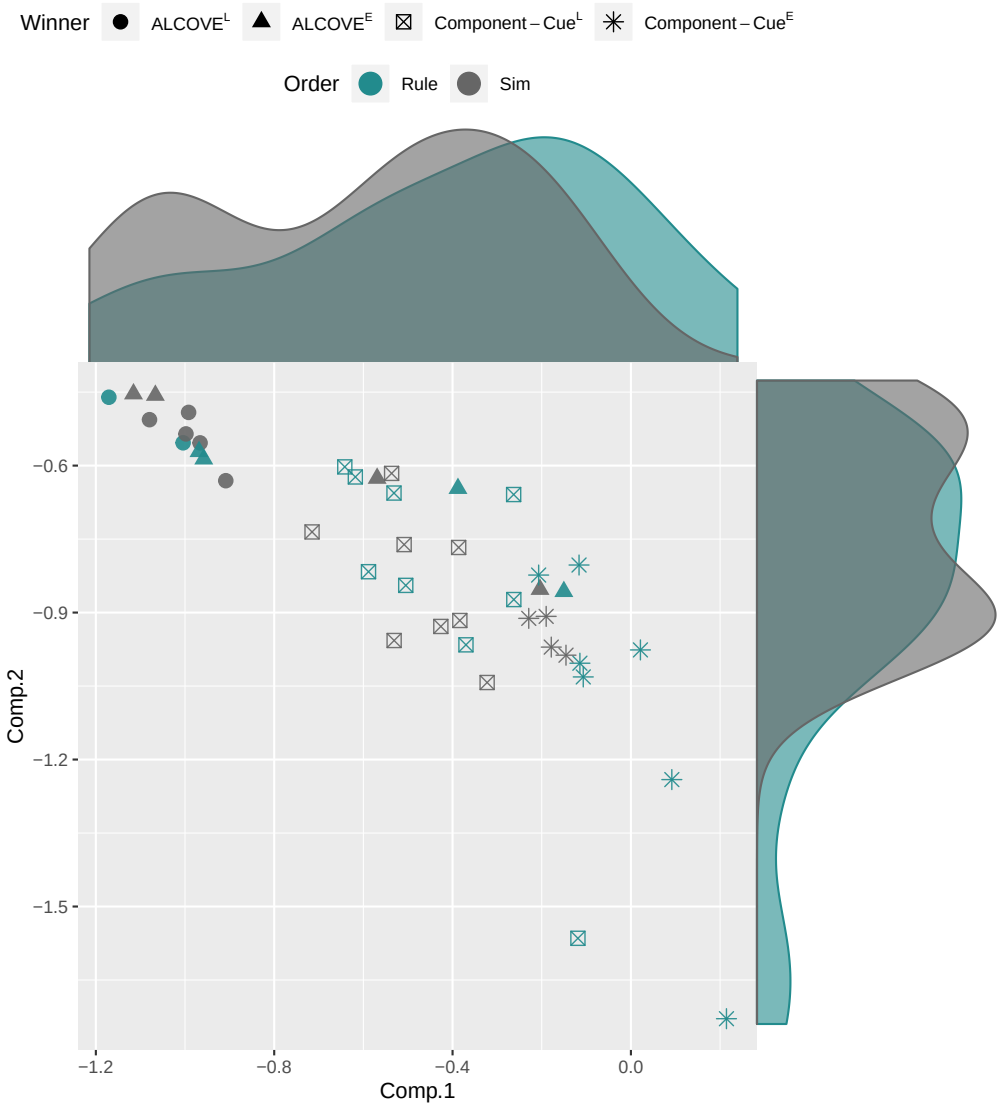


Figure 5.14 – *Result of the investigation of the influence of within-category order on generalization patterns. The graph shows the probability patterns of the winning models projected on the same components of Figure 2.9, as a function of both the winning model (on Experiment I) and the within-category order. In the probability patterns only transfer items were considered. The density functions of the points as a function of the within-category order are also displayed.*

significant (p-value < 0.001). This shows that the generalization patterns of Component-Cue (when Component-Cue is the model that best reproduced participants' responses) were consistent with a rule-based retrieval, confirming that Component-Cue performed better on participants adopting a rule-based strategy.

Conclusion: Firstly, the winning models captured the difference in generalization patterns between rule-based and similarity-based participants that was shown in the analysis of Subsection 2.1.3. Secondly, the analysis confirmed that Component-Cue performed better on participants adopting a rule-based strategy.

Within-Category Order Sensitivity

The aim of this subsection is to determine whether both the learning curves and the generalization patterns of the learning models are sensitive to the within-category order. To this purpose, we averaged across participants the estimated parameters of the models and (using these parameters) we trained the models on three different sequences of stimuli: a random sequence, a rule-based sequence, and a similarity-based sequence.

Figure 5.15a shows the learning curves of the learning models as a function of the stimuli manipulation (Random vs. Rule-Based vs. Similarity-based) and the category membership of the stimuli (A and B). The learning curves of Component-Cue were more sensitive to the within-category order than the learning curves of ALCOVE.

Figure 5.15b shows the projection of the learning curves on the same plan as in Figure 5.1a, as a function of the stimuli manipulation (Random vs. Rule-Based vs. Similarity-based) and the model. Moreover, the projection of the generalization patterns at the end of the training are represented with triangles. Again, the learning curves of Component-Cue were more sensitive to the stimuli manipulation than the learning curves of ALCOVE. Both versions of Component-Cue showed distinct learning curves for different stimuli manipulation. Conversely, both versions of ALCOVE showed distinct learning curves for different stimuli manipulation only in the early stage of learning. Moreover, the analysis of the generalization patterns at the end of the training shows that ALCOVE exhibited similar generalization patterns regardless of the stimuli manipulation. Conversely, Component-Cue exhibited distinct generalization patterns for different stimuli manipulation.

Conclusion: Component-Cue was more sensitive to stimuli manipulation than ALCOVE, showing distinct learning curves and distinct generalization patterns at the end of the training for different stimuli manipulation.

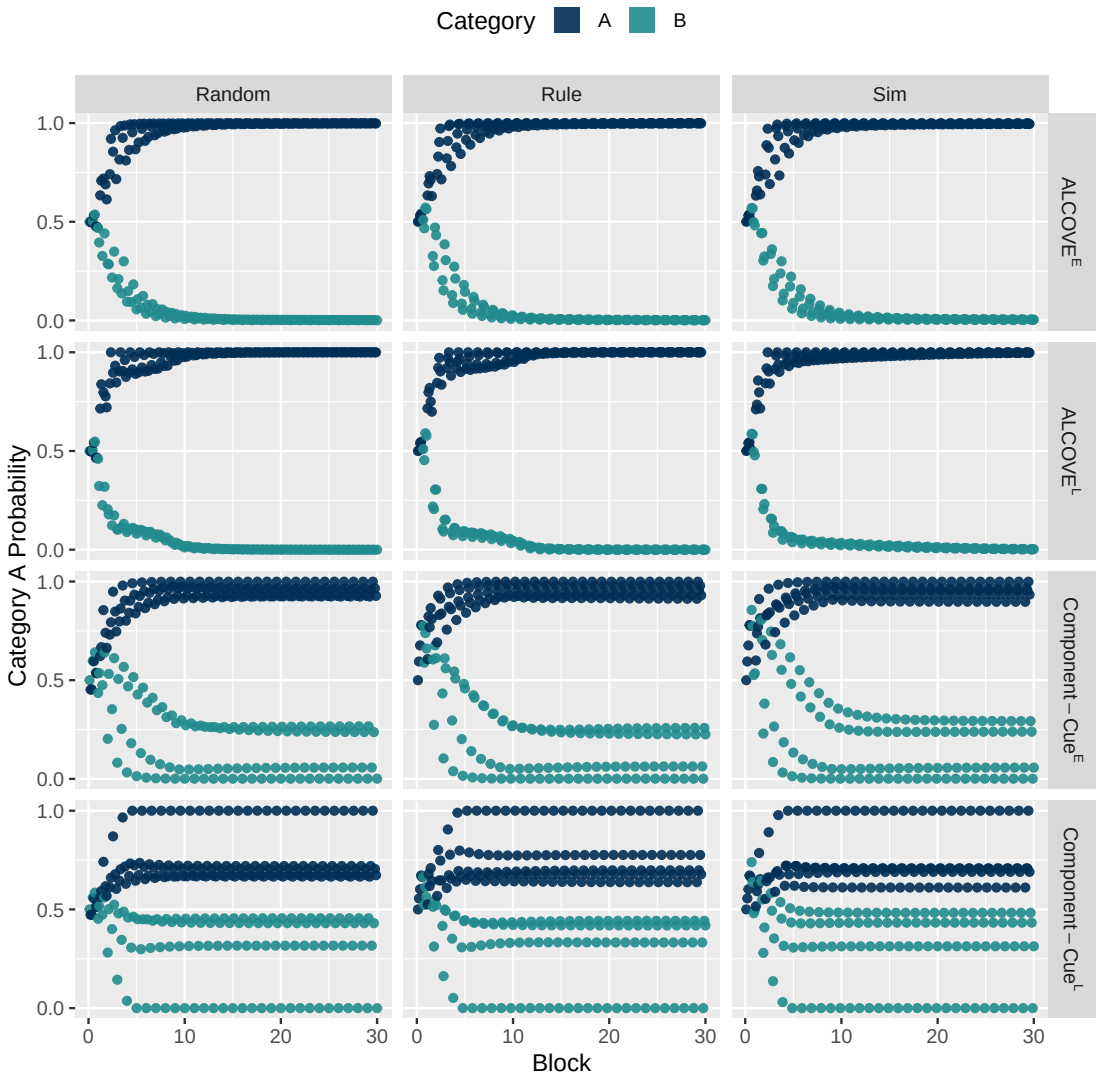


Figure 5.15a – *Learning curves of the learning models as a function of the stimuli manipulation (Random vs. Rule-Based vs. Similarity-based) and the category membership of the stimuli (A and B). The parameters used to train the models are the average (across participants) estimated parameters obtained from the application of the hold-out to Experiment I.*

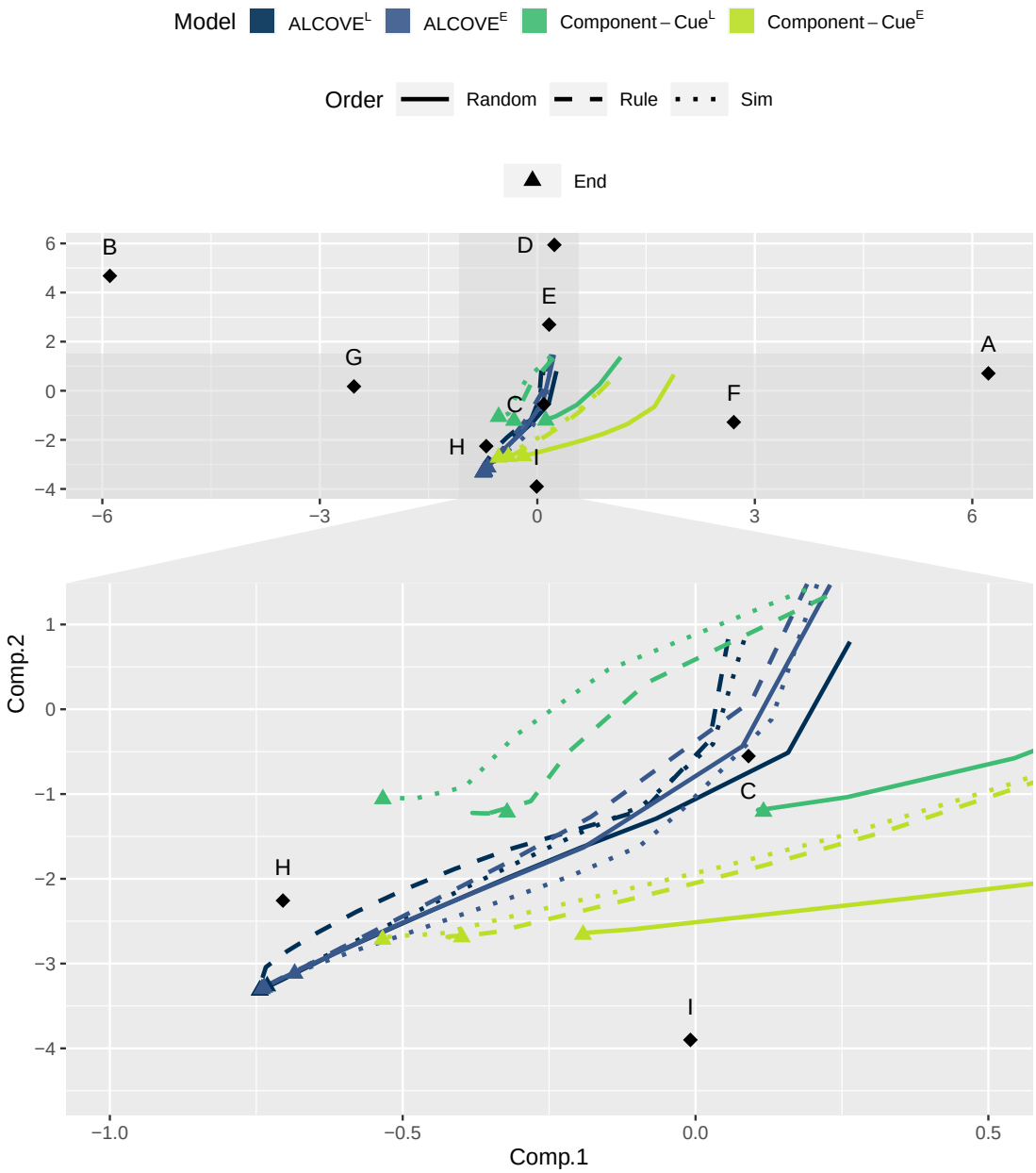


Figure 5.15b – *Projection of the learning curves on the same plan as in Figure 5.1a, as a function of the stimuli manipulation (Random vs. Rule-Based vs. Similarity-based) and the model. The parameters used to train the models are the average (across participants) estimated parameters obtained from the application of the hold-out to Experiment I.*

5.4.4 Rule-Based vs. Similarity-Based from a Model Perspective in Experiment II

The subsection aims to determine whether the within-category order is related to the type of model that best fits Experiment II. Conversely to the previous subsection, only the successful participants were investigated. Indeed, the analysis on generalization patterns was not possible since Experiment II has no transfer phase, the one on learning times was not possible since all observations were censored, and the analysis on the sensibility of the models was less relevant because of the absence of a transfer phase. Additionally, the relation between the number of unsuccessful participants and the type of winning model was investigated.

Number of participants

A Fisher's exact test of independence was performed to determine whether participants in the rule-based order were best fit by Component-Cue, while participants in the similarity-based order were best fit by ALCOVE.

Fisher's exact test of independence. Table 5.3 shows the number of participants whose responses were best predicted by either the Component-Cue models or the ALCOVE models, as a function of both the within-category order (rule-based vs. similarity-based), the evaluation criterion (SSD vs. Likelihood), and the context (Random-Variable vs. Random-Constant vs. Blocked-Constant). No relation between the within-category order and the type of models was visible. The Fisher's exact test of independence confirmed the lack of relation.

Conclusion: No relation was found between the within-category order and the type of winning model.

Number of unsuccessful participants

Here, the possibility that participants' outcome is related to the type of winning model is explored.

	SSD WINNER		LIKELIHOOD WINNER	
	ALCOVE	Component-Cue	ALCOVE	Component-Cue
RANDOM-VARIABLE				
Rule-based	29	5	29	5
Similarity-based	29	5	29	5
RANDOM-CONSTANT				
Rule-based	11	0	11	0
Similarity-based	10	1	10	1
BLOCKED-CONSTANT				
Rule-based	23	0	22	1
Similarity-based	23	0	22	1

Table 5.3 – *Number of participants on Experiment II whose responses were best predicted by either the Component-Cue models or the ALCOVE models, as a function of both the within-category order (rule-based vs. similarity-based), the evaluation criterion (SSD vs. Likelihood), and the context (Random-Variable vs. Random-Constant vs. Blocked-Constant).*

Fisher's exact test of independence. Table 5.4 shows the number of participants whose responses were best predicted by either the Component-Cue models or the ALCOVE models, as a function of participants' outcome (successful vs. unsuccessful), the evaluation criterion (SSD vs. Likelihood), and the context (Random-Variable vs. Random-Constant vs. Blocked-Constant). The Fisher's exact test of independence was only significant for the Random-Variable context (p-value < 0.001 for the SSD criterion and a p-value$=0.004$ for the likelihood criterion), showing that in this context the type of winning model was related to the participants' outcome.

Conclusion: A significant relation between the type of winning model and the participants' outcome was found in the Random-Variable context. No relation was found in the Random-Constant and Blocked-Constant contexts.

	SSD winner		Likelihood winner	
	ALCOVE	Component-Cue	ALCOVE	Component-Cue
Random-Variable				
Successful	38	0	37	1
Unsuccessful	20	10	21	9
Random-Constant				
Successful	20	0	20	0
Unsuccessful	1	1	1	1
Blocked-Constant				
Successful	36	0	36	0
Unsuccessful	10	0	8	2

Table 5.4 – *Number of participants on Experiment II whose responses were best predicted by either the Component-Cue models or the ALCOVE models, as a function of participants' outcome (successful vs. unsuccessful), the evaluation criterion (SSD vs. Likelihood), and the context (Random-Variable vs. Random-Constant vs. Blocked-Constant).*

TO SUM UP *Application of the Advanced Inference Method to Learning Models*

The aim of the chapter is twofold. On one hand to apply the inference method developed in the previous chapter to learning models and determine the model that best describes both Experiment I and II. On the other hand, to investigate whether the distinct network architecture of ALCOVE and Component-Cue is related to the within-category order (rule-based vs. similarity-based).

Visual Representation of Models

The analysis is organized in two parts. The first part aims to analyze the predictions of the models after a period of training, while the second part aims to analyze the learning curves of the models. The first analysis on the predictions of the models showed that the learning models are nested, with the Component-Cue models included in the ALCOVE models and the linear versions included in the exponential versions. The second analysis on the learning curves showed that high values of λ_ω and λ_w produce learning curves with a high variability; high values of c lead to learning curves with a lower variability and a higher probability of correctly classifying the stimuli (as compared to low values of c); high values of ϕ amplify the value of the classification probability; and high value of b shrink the space of the classification probability to a limited area centered around 0.5.

Parameter Estimation

A study aiming to validate the consistency of the MLE on simulated data was conducted. The result of the analysis showed that the classification probability was accurately estimated when the size of the dataset was equal to or greater than 40 blocks. Conversely, to accurately estimate the parameters of the models a higher number of observations is needed (80-160 blocks depending on the parameter).

Model Selection

Learning models were compared using the hold-out method, the simplest kind of cross-validation method. A preliminary validation of the identifiability of the

learning models via the hold-out method was necessary. The result varied among experiments and models. In Experiment I, the model that has the lowest evaluation is the generative model with a probability of 88% when it is Component-Cue^L, with a probability of 65% when it is Component-Cue^E, and with a probability of 58% when it is either ALCOVE^L or ALCOVE^E. However, when only the type of the model is considered regardless of the version, the probability that the model with the lowest evaluation is the generative model raised to 85% (on average). In Experiment II, the model that has the lowest evaluation is the generative model with a probability of 76% (on average) when it is either Component-Cue^L or Component-Cue^E, and with a probability of 52% (on average) when it is either ALCOVE^L or ALCOVE^E. Again, when only the type of the model is considered regardless of the version, the probability that the model with the lowest evaluation is the generative model raised to 95% for the Component-Cue models and to 76% for the ALCOVE models.

Experimental Data Analysis

The section is organized in four parts. The first two are devoted to the application of the hold-out method to both Experiment I and II. The last two are devoted to the investigation of the relation between the within-category presentation order (rule-based vs. similarity-based) and the type of model that best describes the experimental data (ALCOVE vs. Component-Cue). In both experiments, the hold-out method was applied to each participant, separately. In Experiment I, the learning phase was used as the training set and the transfer phase as the testing set. In Experiment II, the early 80% of the learning phase was used as the training set and the last 20% as the testing set. The results showed that the majority of the participants in Experiment I were best fit by Component-Cue (one of the two versions). Conversely, almost all participants in Experiment II were best fit by ALCOVE (one of the two versions). The relation between the within-category presentation order and the type of model that best describes the experimental data was investigated through the following analyses:

Number of participants with the Fisher's exact test of independence (not significant). This analysis aimed to determine whether the number of participants whose responses were best predicted by a specific type of model (ALCOVE vs. Component-Cue) was related to the within-category order. The Fisher's

exact test of independence was not significant in both Experiment I and Experiment II.

Learning times with the Cox proportional-hazards model (not significant). This analysis was only performed on Experiment I and investigated whether participants whose responses were best predicted by Component-Cue had better chances to reach the learning criterion as compared to participants whose responses were best predicted by ALCOVE. The Cox model was not significant.

Generalization patterns with PCA and Wilcoxon-Mann-Whitney test (significant). In the same vain as in Subsection 2.1.3, the generalization patterns of the winning models were examined as a function of the within-category order. The results showed a significant difference in location between the predictions of the winning models of rule-based participants and the predictions of the winning models of similarity-based participants. Moreover, a significant difference in location between the predictions of the Component-Cue winners and the predictions of the ALCOVE winners was found. Therefore, *i)* the models that best fit participants' performance captured the difference in generalization patterns between rule-based and similarity-based participants, and *ii)* Component-Cue better reproduces participants adopting a rule-based strategy.

Within-category order sensitivity. This analysis was only performed on Experiment I and aimed to determine whether both the learning curves and the generalization patterns of the learning models are sensitive to the within-category order. The results showed that Component-Cue was more sensitive to stimuli manipulation than ALCOVE. Indeed, it exhibited distinct learning curves and distinct generalization patterns at the end of the training for different stimuli manipulations. Conversely ALCOVE showed distinct learning curves for different stimuli manipulations only in the early stage of learning.

Number of unsuccessful participants with the Fisher's exact test of independence (significant in the Random-Variable context of Experiment II). This analysis was only performed on Experiment II and explored whether the type of winning model (ALCOVE vs. Component-Cue) was related to the participants' outcome (successful vs. unsuccessful). The Fisher's exact test of independence showed a significant relation between type of winning model and participants' outcome only in the Random-Variable context.

6

Alternative Inference Method for Learning Data

Contents

We mentioned in several occasions that transfer models are not able to evolve over time (see "Ability to Learn" in Chapter 1 and Section 3.2). The lack of temporal dynamic prevents transfer models to achieve good quantitative predictions during learning. However, the following question raises: Are there statistical methods that would allow transfer models to accurately reproduce learning? The aim of the present chapter is to investigate two of them: the segmentation and the segmentation/clustering.

Outline of this chapter

Firstly, we describe the segmentation technique and apply it to the GCM. Secondly, we describe the segmentation/clustering technique and, again, apply it to the GCM.

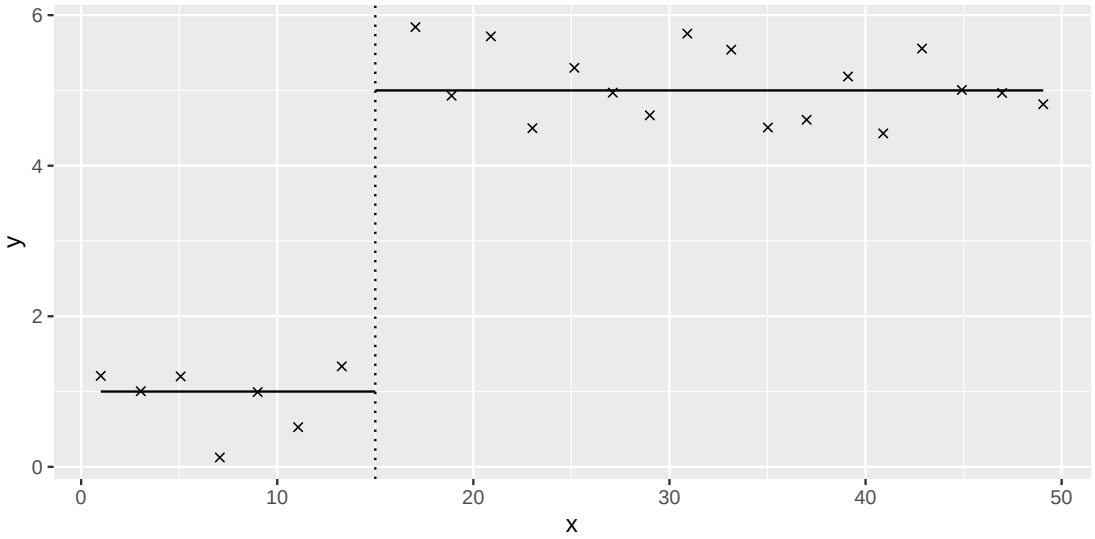


Figure 6.1 – *Example of segmentation method. The x marks represents the time series and the dotted line represents the change-point of the series. The points belonging to the two segments in which the data were split can be summarized with two horizontal lines (the intercept of the line is equal to the average y value). The segmentation aims to find the number and the positions of the change-points.*

6.1 Segmentation Method with Transfer Models

Among the methods that would allow transfer models to reproduce learning, one of the firsts that comes to mind is to partition the learning process into multiple segments and estimate the parameters of the model on these segments, separately. In a figurative way, this would correspond to consider a step function instead of a constant function to reproduce a (strictly) increasing or decreasing function. There are multiple techniques that allow the effective partition of the learning process into segments. One of them is the (off-line) segmentation method. The aim of this method is to detect abrupt changes (also called change-points) affecting the parameters of the model. Before describing the mathematical framework, let us give a basic example.

Example 6.1. Let us consider the time series illustrated in Figure 6.1 (the x marks). This series is characterized by a change-point (the dotted line). More specifically, there is an abrupt change in the statistical behavior of the points. Thus, the change-point partitions the points into two segments, each of them grouping points with similar statistical properties. Since the points that belong to these two segments have similar characteristics, they can be summarized with two horizontal lines (the intercept of the

line is equal to the average y value). The aim of the segmentation method is to detect the number (1 in this case) and the positions of the change-points (i.e., the dotted line) that occur in the time series. Moreover, the segmentation allows the reduction of the information included in the time series to the position of the change-point(s) and to the "summary" of the segments (i.e., the horizontal lines). ⊠

6.1.1 Mathematical Framework

The description of the segmentation is given when the number of change-points is fixed. Considerations about the selection of the number of change-points are given at the end of the subsection.

Model. Let $y_1, \ldots, y_n$ be n observations and let $Y_1, \ldots, Y_n$ be n random variables such that y_i is a realization of Y_i (for $i = 1, \ldots, n$). Let us assume that the process $Y_1, \ldots, Y_n$ is affected by K abrupt changes at unknown coordinates $\tau = \{\tau_1, \ldots, \tau_K\}$ (with the convention $\tau_0 = 1$ and $\tau_{K+1} = n + 1$). The K change-points define a partition of the observations into $K + 1$ segments $S_1, \ldots, S_{K+1}$ such that:

$$S_k = \{y_t,\, t \in [\tau_{k-1}, \tau_k)\}\,.$$

According to the segmentation model, the random variables have the following distribution:

$$Y_t \sim f(\theta_k) \quad \forall t \in S_k,$$

where the parameter θ_k can assume an infinite number of values. In our case, the function f assumes the following form:

$$f(\theta_k) = \mathcal{B}\left(\mathbb{P}_M^{\theta_k}\left(A \,|\, x^{(t)}\right)\right),$$

where M is a transfer model and $x^{(t)}$ is the t-th stimulus (it is associated to the observation y_t).

Goal. The goal of the segmentation method is to infer from the observed data the positions of the change-points. More specifically, given the observed data $y_1, \ldots, y_n$, the

aim is to find $\tau = \{\tau_1, \ldots, \tau_K\}$ such that the cost of splitting the observed data into $K+1$ segments is minimal:

$$\min_{\tau_1,\ldots,\tau_K} \sum_{k=1}^{K+1} \mathcal{C}_{\tau_{k-1}:\tau_k}.$$

The quantity $\mathcal{C}_{\tau_{k-1}:\tau_k}$ represents the cost of the k-th segment and (in our case) is given by the likelihood of the model evaluated on the segment. In more detail, the cost of the k-th segment is defined as follows:

$$\mathcal{C}_{\tau_{k-1}:\tau_k} = \min_{\theta} \left\{ \sum_{j \in [\tau_{k-1}, \tau_k)} -\log \mathbb{P}_M^{\theta} \left(A \,|\, x^{(j)} \right) \right\}. \tag{6.1}$$

Algorithm. To recover the segmentation of minimal cost with K change-points, the standard dynamic programming algorithm [AL89] is used (it is also called the segment neighborhood algorithm). This algorithm is based on the Bellman's principle of optimality [BD62] according to which, if a segmentation is optimal, then any sub-segmentation of this segmentation is also optimal. Mathematically, this principle can be expressed by the following update rule:

$$\mathfrak{C}_{1:t}^{k} = \min_{\tau < t} \left\{ \mathfrak{C}_{1:\tau}^{k-1} + \mathcal{C}_{\tau:t} \right\}, \tag{6.2}$$

where $\mathfrak{C}_{1:t}^{k}$ is the cost to partition the segment $\{y_s,\, s \in [1, t)\}$ into $k+1$ segments and k is a constant. The iteration of the previous update rule allows us to recover the position of the change-points $\tau_1, \ldots, \tau_K$ as well as the values $\theta_1, \ldots, \theta_{K+1}$. The overall time complexity of this algorithm is $O(Kn^2)$. Indeed, the time complexity of the update rule (Equation 6.2) is $O(t)$ and, in order to recover $\mathfrak{C}_{1:n+1}^{K}$, the update rule has to be applied for every t smaller than $n+1$ and every k smaller than K, which makes $O(Kn^2)$.

Choice of the number of change-points. To estimate the number of change-points the method proposed by Lavielle in [Lav05] was implemented. The technique can be summarized as follows: *i)* to examine the way the segmentation cost decreases as the number of change-points increases, and *ii)* to determine the number of change-points with which the segmentation cost ceases to decrease significantly. In other words, the number of optimal change-points is found by looking for a break in the slope of the cost function (an example of its application is given in the next subsection).

6.1.2 Application to the Generalized Context Model (GCM)

The application of the segmentation method was limited to the simplest model among the selected transfer models: the GCM. Moreover, the segmentation was only performed on the learning phase of the Random-Constant context of Experiment II because of time constraints. The Random-Constant context was preferred to the other experiments because of its limited number of participants (22), which would facilitate the visualization of the segmentation.

Technical Aspects

i. The observations consisted in the learning phase of the Random-Constant context of Experiment II. The segmentation method was only applied to the GCM.

ii. The likelihood was only expressed as a function of the sensitive parameter c. The attention-weight parameters were fixed and the attention was equally shared between the dimensions.

iii. The number of change-points was fixed at 1 for every participant. The choice $K = 1$ was the result of the implementation of the method proposed by Lavielle. Figure 6.2 (on the top) shows its application on the observations of the Random-Constant context of Experiment II. The graph shows the evolution of the segmentation cost for each participant, as a function of the number of change-points. A break in the slope of the segmentation cost is detected when the number of change-points is equal to 1.

Results

Figure 6.2 (on the bottom) shows the application of the segmentation with exactly 1 change-point to the learning phase of the Random-Constant context of Experiment II. All participants improved their performance over time and almost every participant began the classification task with a low value of the sensitive parameter c (i.e., low performance/almost random classification) and finished the classification task with a high value of the sensitive parameter c (i.e., high/perfect performance). The two participants with a final sensitive parameter that was smaller than 10 are the two unsuccessful participants.

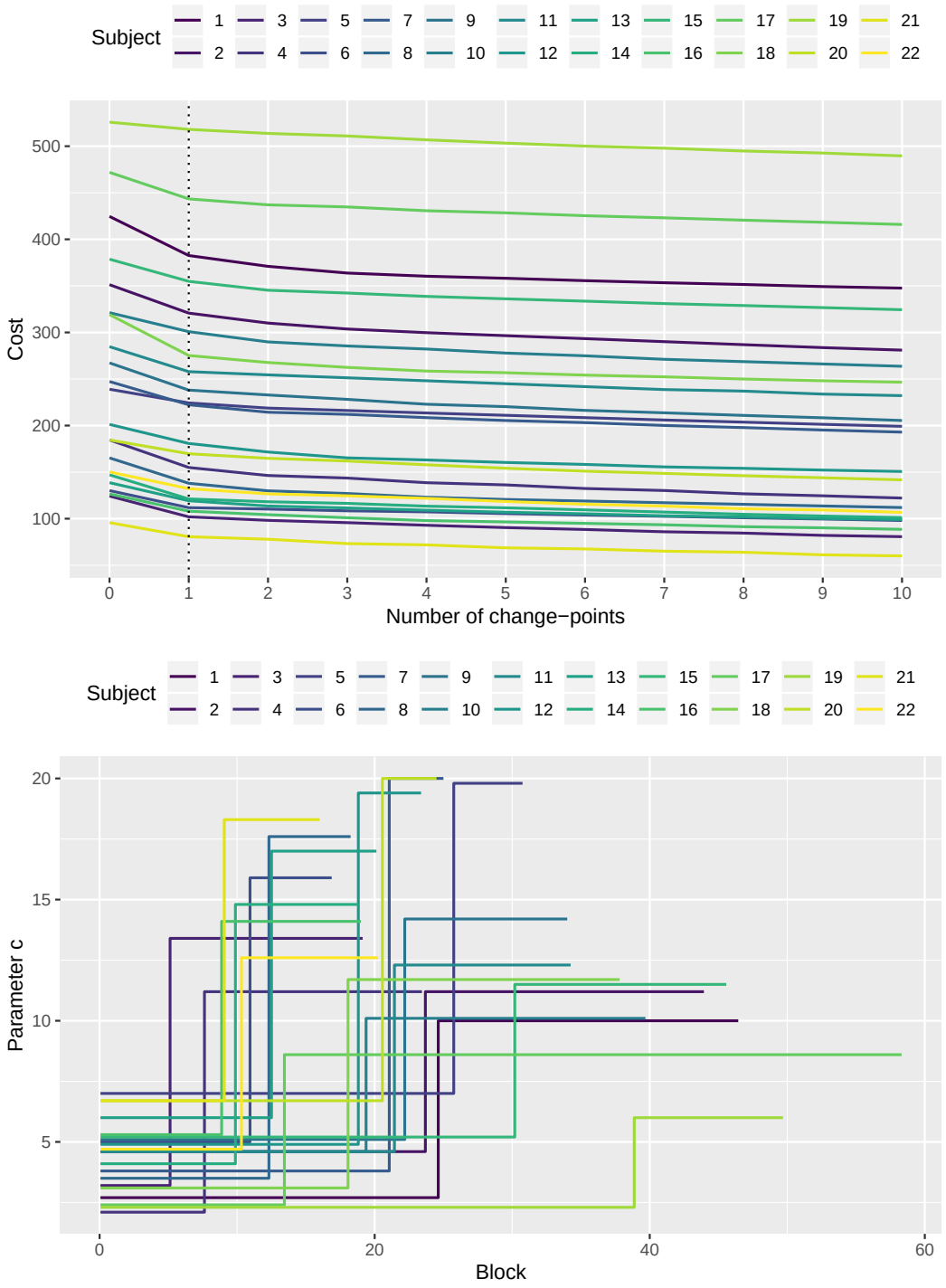


Figure 6.2 – *Application of the segmentation method to the Random-Constant context of Experiment II. The method was separately applied on the learning phase of each participant. On the top, the segmentation cost of each participant as a function of the number of change-points. A break in the slope of the segmentation cost is found with 1 change-point. On the bottom, the result of the segmentation with exactly 1 change-point.*

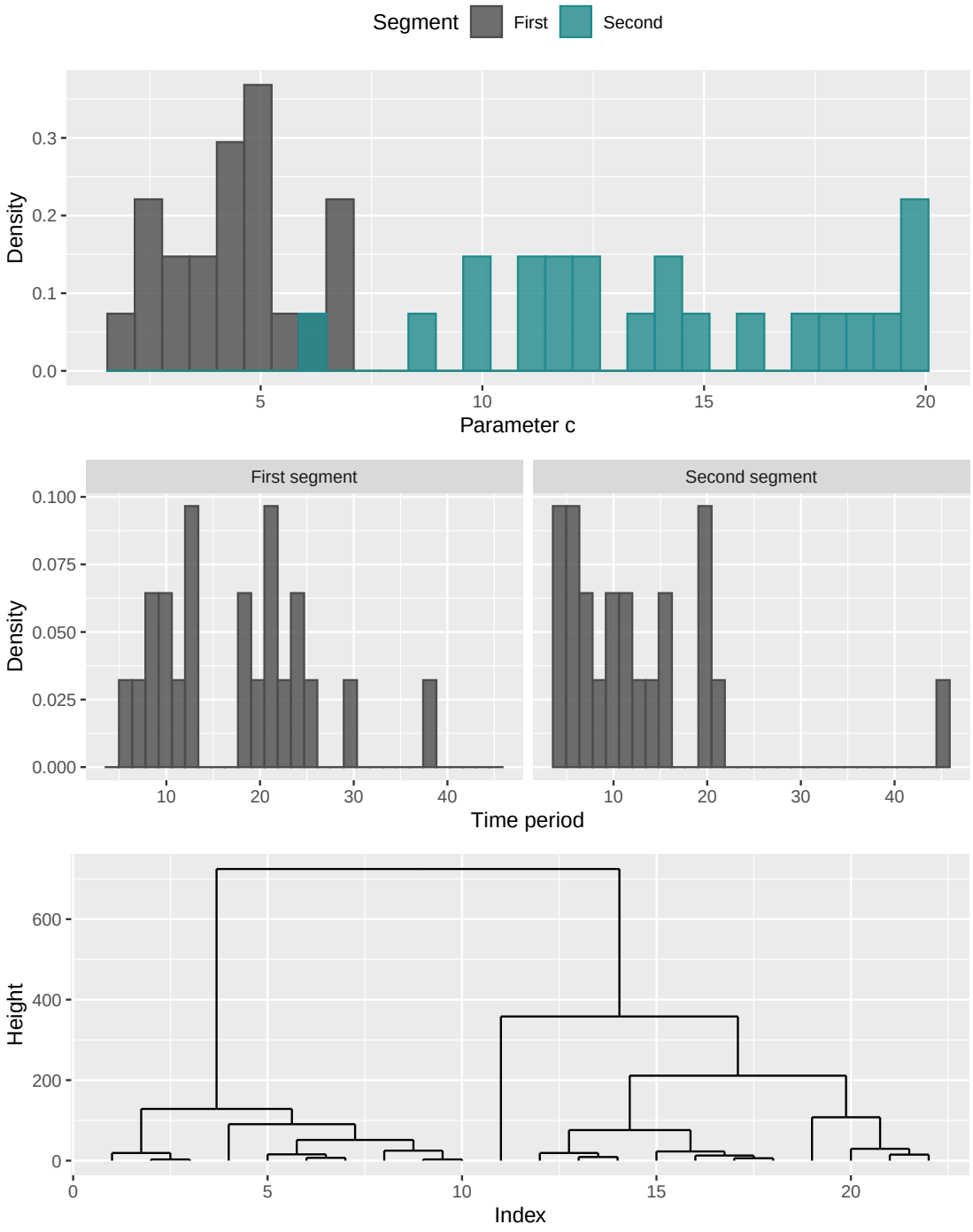


Figure 6.3 – *Analysis of the results obtained from the application of the segmentation with exactly 1 change-point to the Random-Constant context of Experiment II. On the top, the variability of the estimated sensitive parameter, as a function of the segment (i.e., first or second). On the middle, the variability of the time period spent by participants on each segment (i.e., first or second). On the bottom, the dendrogram applied to the time period spent by participants on the first segment (two clusters are identified).*

In Figure 6.3, the results of the segmentation are analyzed. The graph on the top shows the variability of the estimated sensitive parameter, as a function of the segments (i.e., first or second). The average sensitive parameter on the first segment was lower than the average sensitive parameter on the second segment (4.4 vs. 14). This confirms the previous affirmations regarding Figure 6.2 (on the bottom).

Figure 6.3 (on the middle) shows the time period spent by participants on each segment. On the first segment, two clusters of participants are detected using a dendrogram (see Figure 6.3 on the bottom): those who spent on average 10 blocks on the first segment and those who spent on average 22 blocks on the first segment. Since the first segment is associated with a low sensitive parameter (i.e., a low performance) and since all participants improved their performance over time (i.e., their performance on the second segment was better), the two clusters correspond to two groups of participants with different learning speed (i.e., high and low learning speed). The participants who spent less than 15 blocks on the first segment are called high-speed participants, while those who spent at least 15 blocks on the first segment are called low-speed participants.

A Fisher's exact test of independence was performed to determine whether the two clusters (high-speed participants vs. low-speed participants) were related to the within-category order (rule-based vs. similarity-based). The test was applied to Table 6.1, in which the number of high-speed and low-speed participants is shown as a function of the within-category order. The test was not significant (p-value=0.67), showing no relation between participants' speed and within-category order.

	High-speed	Low-speed
Rule-based	6	5
Similarity-based	4	7

Table 6.1 – *Number of high-speed and low-speed participants in the first segment, as a function of the within-category order (rule-based vs. similarity-based).*

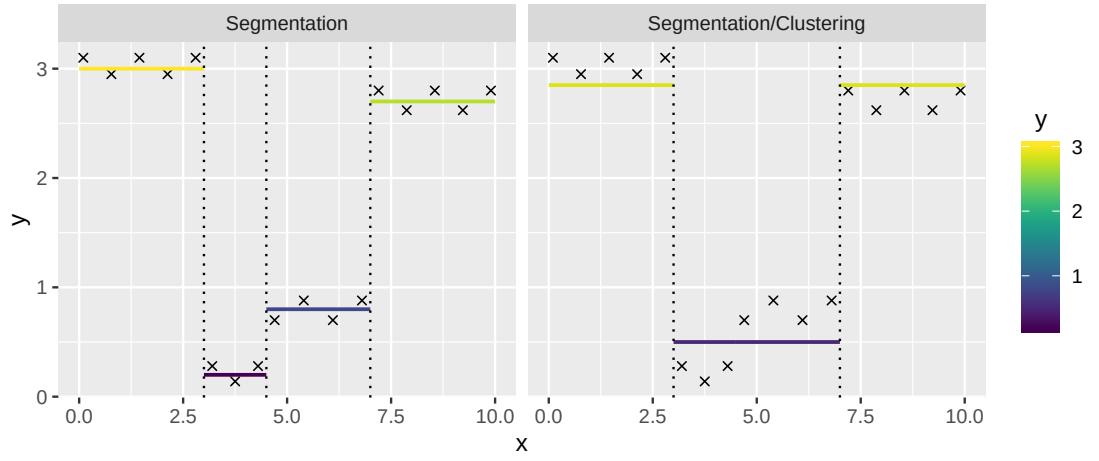


Figure 6.4 – *Illustration of the segmentation/clustering method as compared to the segmentation one. On the left, the segmentation technique detects the optimal change-points to partition the data into segments. On the right, the segmentation/clustering technique organizes the segments into a finite number of clusters. For instance, the second and the third segments as well as the first and the fourth are grouped into the same cluster because of their similar values.*

6.2 Segmentation/Clustering Method with Transfer Models

The segmentation allowed us to partition the participants' observations into a finite number of segments, each of them grouping observations with similar characteristics (see Figure 6.4, on the left). Although it represents an effective solution to apply transfer models to learning data, its application raises some issues. Firstly, the number of observations per segment could be too small to accurate estimate the parameters. Secondly, the segmentation method does not allow one to easily compare the participants' learning dynamic.

The aim of the segmentation/clustering method is to address these issues. The segmentation/clustering method assumes that *i)* the level of each segment (in our case the value of the sensitive parameter) can only take a limited number of values, and *ii)* each level is associated with a cluster. The principle of the segmentation/clustering method is illustrated in Figure 6.4, on the right. In this example, the second and third segments as well as the first and the fourth segments are grouped into the same cluster because of their similar values.

6.2.1 Mathematical Framework

The description of the segmentation/clustering model is given when the number of change-points and the number of clusters is fixed. The selection of the number of change-points and clusters is addressed in the last paragraph.

Model. Let $s \in \mathcal{S}$ be a participant and let $y_1^s, \dots, y_{n_s}^s$ be his n_s observations. Let $Y_1^s, \dots, Y_{n_s}^s$ be n_s random variables such that y_i^s is a realization of Y_i^s ($i = 1, \dots, n_s$). Let us assume that the process $Y_1^s, \dots, Y_n^s$ is affected by K abrupt changes at unknown coordinates $\tau^s = \{\tau_1^s, \dots, \tau_K^s\}$ (with the convention $\tau_0^s = 1$ and $\tau_{K+1}^s = n + 1$). The K change-points define a partition of the observations into $K + 1$ segments $S_1^s, \dots, S_{K+1}^s$ such that:

$$S_k^s = \{y_t^s,\, t \in [\tau_{k-1}^s, \tau_k^s)\}\,.$$

According to the segmentation/clustering model (applied to our case), the random variables have the following distribution:

$$Y_t^s \sim \mathcal{B}\left(\mathbb{P}_M^{\theta_k}\left(A \,|\, x^{(t),s}\right)\right) \qquad \forall t \in S_k^s,$$

where the parameter θ_k can only take P values (i.e., $\theta_k \in \{\vartheta_1, \dots, \vartheta_P\}$), M is a transfer model, and $x^{(t),s}$ is the t-th stimulus presented to participant s (it is associated to the observation y_t^s). The quantity P denotes the number of clusters, while $\vartheta_1, \dots, \vartheta_P$ are the values associated to each cluster. This means that, in addition to the spatial organization of the data in segments, a secondary organization of the segments in clusters is considered (the clusters are the same for all participants). In our case, the clusters code different learning performance (e.g., random classification, perfect classification, etc.) For a deeper mathematical description of this model, we refer to [Pic+07].

Goal. The goal of the segmentation/clustering method is to infer from the observed data the positions of the change-points and the values associated to the clusters. More specifically, given the observed data $y_1^s, \dots, y_{n_s}^s$ (for every $s \in \mathcal{S}$), the aim is to find $\tau^s = \{\tau_1^s, \dots, \tau_K^s\}$ and $\vartheta = \{\vartheta_1, \dots, \vartheta_P\}$ such that the cost of the segmentation is minimal:

$$\sum_{s \in \mathcal{S}} \min_{\tau_1^s, \dots, \tau_K^s} \sum_{k=1}^{K+1} \mathcal{C}_{\tau_{k-1}^s : \tau_k^s}.$$

The quantity $\mathcal{C}_{\tau_{k-1}^s:\tau_k^s}$ represents the cost of the k-th segment of participant s and (in our case) is given by the likelihood of the model evaluated on the segment. In more detail, the cost of the k-th segment of participant s is defined as follows:

$$\sum_{s\in\mathcal{S}} \min_{\tau_1^s,\dots,\tau_K^s} \sum_{k=1}^{K+1} \min_{\theta\in\{\vartheta_1,\dots,\vartheta_P\}} \left\{ \sum_{j\in[\tau_{k-1}^s,\tau_k^s)} -\log \mathbb{P}_M^\theta \left(A \mid x^{(j)}\right) \right\}. \tag{6.3}$$

The parameter θ in Equation 6.3 can only take a limited number of values (i.e., $\vartheta_1,\dots,\vartheta_P$). Conversely, in Equation 6.1 the parameter θ could take an infinite number of values.

Algorithm. To apply the segmentation/clustering model we used a hybrid algorithm called dynamic programming-expectation maximization (DP-EM) [Pic+07]. Since the segmentation/clustering model combines segmentation and mixture models, the hybrid algorithm combines dynamic programming (DP) algorithm, used with segmentation models, and expectation maximization (EM) algorithm, used with mixture models. The principle of the DP-EM is the following: when the values $\vartheta = \{\vartheta_1,\dots,\vartheta_P\}$ are known, the position of the change-points $\tau^s = \{\tau_1^s,\dots,\tau_K^s\}$ is computed using the DP algorithm (for each $s\in\mathcal{S}$), and once the change-point coordinates $\tau^s = \{\tau_1^s,\dots,\tau_K^s\}$ are estimated (for each $s\in\mathcal{S}$), the EM algorithm is used to estimate (again) the values $\vartheta = \{\vartheta_1,\dots,\vartheta_P\}$. The DP-EM algorithm is composed of the following steps (we recall that the algorithm is performed for a fixed number of segments K and a fixed number of clusters P):

Step 0. Let $\mathcal{P}$ denote the set of P clusters. The step zero consists in associating a value ϑ_p to each cluster $p\in\mathcal{P}$.

Step 1. Given the values $\vartheta = \{\vartheta_1,\dots,\vartheta_P\}$, the first step consists in finding the change-point coordinates $\tau^s = \{\tau_1^s,\dots,\tau_K^s\}$ for each participant $s\in\mathcal{S}$ such that the cost of the segmentation of each participant is minimal (Equation 6.3 without the external sum across the participants). This step is performed by means of the DP algorithm.

Step 2. The second step consists in considering all of the segments associated to a specific cluster p (among all participants) and recomputing the value ϑ_p such that the new value minimizes the cost of the group of segments associated to cluster p:

$$\vartheta_p \in \arg\min_\theta \sum_{j:p_{y_j^s}=p} -\log \mathbb{P}_M^\theta \left(A \mid x^{(j)}\right) \qquad \forall p\in\mathcal{P}.$$

The term $p_{y_j^s}$ represents the cluster to which point y_j^s is associated.

Step 3. The third step consists in iterating the algorithm, returning to Step 1. The reader is referred to [Pic+07] for the proof of the convergence properties of the hybrid algorithm.

Choice of the number of change-points and clusters. In the current study, we decided to fix the number of change-points and clusters in order to simplify the implementation of the algorithm and the interpretability. However, we plan to use the same model selection technique found in [Pic+07] in future work.

6.2.2 Application to the Generalized Context Model (GCM)

Again, the application of the segmentation/clustering method was limited to the simplest model among the selected transfer models: the GCM. Moreover, the segmentation was only performed on the learning phase of the Random-Constant context of Experiment II because of time constraints. The segmentation/clustering method was performed two times: the first time with 1 change-point and 3 clusters, and the second time with 2 change-points and 3 clusters. The choice of 3 clusters was motivated by the wish to consider 3 learning regimes: a low/random classification performance, a middle classification performance, and a high/perfect classification performance. The aim of the application of the segmentation/clustering method with 1 change-point and 3 clusters was to analyze the evolution of the participants' learning regimes (e.g., evolution from a low/random classification performance to a high/perfect classification performance). Conversely, the aim of the application of the segmentation/clustering method with 2 change-point and 3 clusters was to detect clusters of high-speed and low-speed participants, assuming that all participants moved through the same learning regimes.

Technical Aspects

i. The observations consisted in the learning phase of the Random-Constant context of Experiment II. The segmentation method was only applied to the GCM.

ii. The likelihood was only expressed as a function of the sensitive parameter c. The attention-weight parameters were fixed and the attention was equally shared between the dimensions.

iii. The segmentation/clustering method was performed two times: the first time with $K = 1$ and $P = 3$, and the second time with $K = 2$ and $P = 3$.

iv. The algorithm was iterated 5 times.

Results

Case with 1 change-point and 3 clusters. The final values of the sensitive parameter c for the three clusters were 2.8, 6, and 12.8. Therefore, the clusters were associated with the following learning regimes: a low/random classification performance (when $c = 2.8$), a middle classification performance (when $c = 6$), and a high/perfect classification performance (when $c = 12.8$). The participants moved only through the following regimes: from middle to high classification performance, from low to high classification performance, and from low to middle classification performance.

Figure 6.5 (on the top) shows the number and percentage of participants, as a function of the evolution of their performance (i.e., from middle to high vs from low to high vs. from low to middle) and the within-category order (i.e., rule-based vs. similarity-based). Most of the participants moved from a middle to a high classification performance. Moreover, the majority of participants in the rule-based order moved from a middle to a high classification performance, while the evolution of the participants' performance in the similarity-based order was varied. A Fisher's exact test of independence was performed to determine whether the evolution of participants' performance was related to the within-category order. The test applied to Table 6.2 (on the top) fell short of significance (p-value=0.1), showing no relation between the two variables.

Figure 6.5 (on the middle) shows the time period spent by participants on each segment. On the first segment, three clusters of participants are detected using a dendrogram (see Figure 6.5 on the bottom): those who spent on average 10 blocks on the first segment, those who spent on average 23 blocks on the first segment, and those who spent on average 35 blocks on the first segment. Since the first segment is associated with a low/middle sensitive parameter and since all participants improved their performance over time, the three clusters correspond to three groups of participants with different learning speed (i.e., high, middle, and low learning speed).

A Fisher's exact test of independence was performed to determine whether the three clusters (high-speed participants vs. middle-speed participants vs. low-speed participants)

	Middle-High	Low-High	Low-Middle
Rule-based	9	2	0
Similarity-based	4	5	2

	High-speed	Middle-speed	Low-speed
Rule-based	6	4	1
Similarity-based	4	5	2

Table 6.2 – *Analysis of the application of the segmentation/clustering to the Random-Constant context of Experiment I with 1 change-point and 3 clusters. On the top, the number of participants, as a function of both the evolution of their performance (i.e., from middle to high vs from low to high vs. from low to middle) and the within-category order (i.e., rule-based vs. similarity-based). On the bottom, the number of high-speed, middle-speed, and low-speed participants in the first segment, as a function of the within-category order.*

were related to the within-category order (rule-based vs. similarity-based). The test was applied to Table 6.2 (on the bottom) and was not significant (p-value=0.7).

Case with 2 change-points and 3 clusters. The final values of the sensitive parameter c for the three clusters were 2.1, 7.1, and 15.5. Thus, again, the clusters were associated with the following learning regimes: a low/random classification performance (when $c = 2.1$), a middle classification performance (when $c = 7.1$), and a high/perfect classification performance (when $c = 15.5$). Almost all participants moved through the following regimes (19 over 22 participants): from low to middle to high classification performance. In what follows, we only analyze the data of the participants that moved through the three learning regimes (in an ascending order).

Figure 6.6 (on the top) shows the time period spent by participants on each segment. On both the first and second segments, two clusters of participants are detected using a dendrogram (see Figure 6.5 on the middle for the first segment and on the bottom for the second segment). On the first segment, the high-speed participants spent less than 10 blocks on the segment, while low-speed participants spent at least 10 blocks on the segment. On the second segment, the high-speed participants spent less than 20 blocks on the segment, while low-speed participants spent at least 20 blocks on the segment.

	High-speed	Low-speed
Rule-based	11	0
Similarity-based	6	5

	High-speed	Low-speed
Rule-based	9	2
Similarity-based	7	4

Table 6.3 – *Analysis of the application of the segmentation/clustering to the Random-Constant context of Experiment I with 2 change-point and 3 clusters. The number of high- and low-speed participants in the first (on the top) and second (on the bottom) segment, as a function of the within-category order (rule-based vs. similarity-based).*

A Fisher's exact test of independence was performed to both segments determine whether the two clusters (high-speed participants vs. low-speed participants) were related to the within-category order (rule-based vs. similarity-based). The test was applied to both of the tables in Table 6.3 and was only significant on the first segment (p-value=0.035 on the first segment and p-value=0.63 on the second segment).

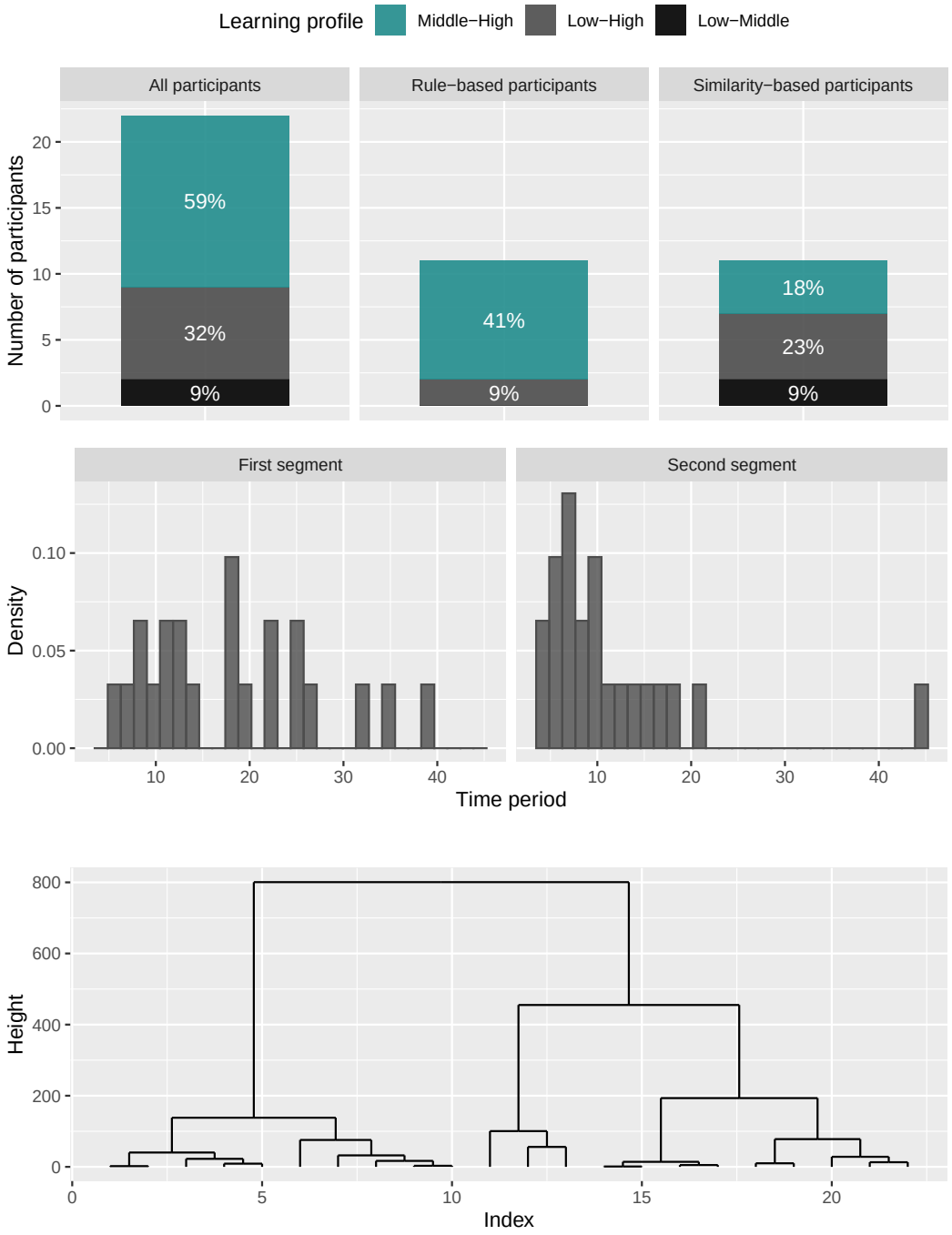


Figure 6.5 – *Analysis of the application of the segmentation/clustering to the Random-Constant context of Experiment I with 1 change-point and 3 clusters. On the top, the number and percentage of participants, as a function of the evolution of their performance (i.e., from middle to high vs from low to high vs. from low to middle) and the within-category order (i.e., rule-based vs. similarity-based). On the middle, the time period spent by participants on each segment. On the bottom, the dendrogram of the time period spent by participants on the first segment.*

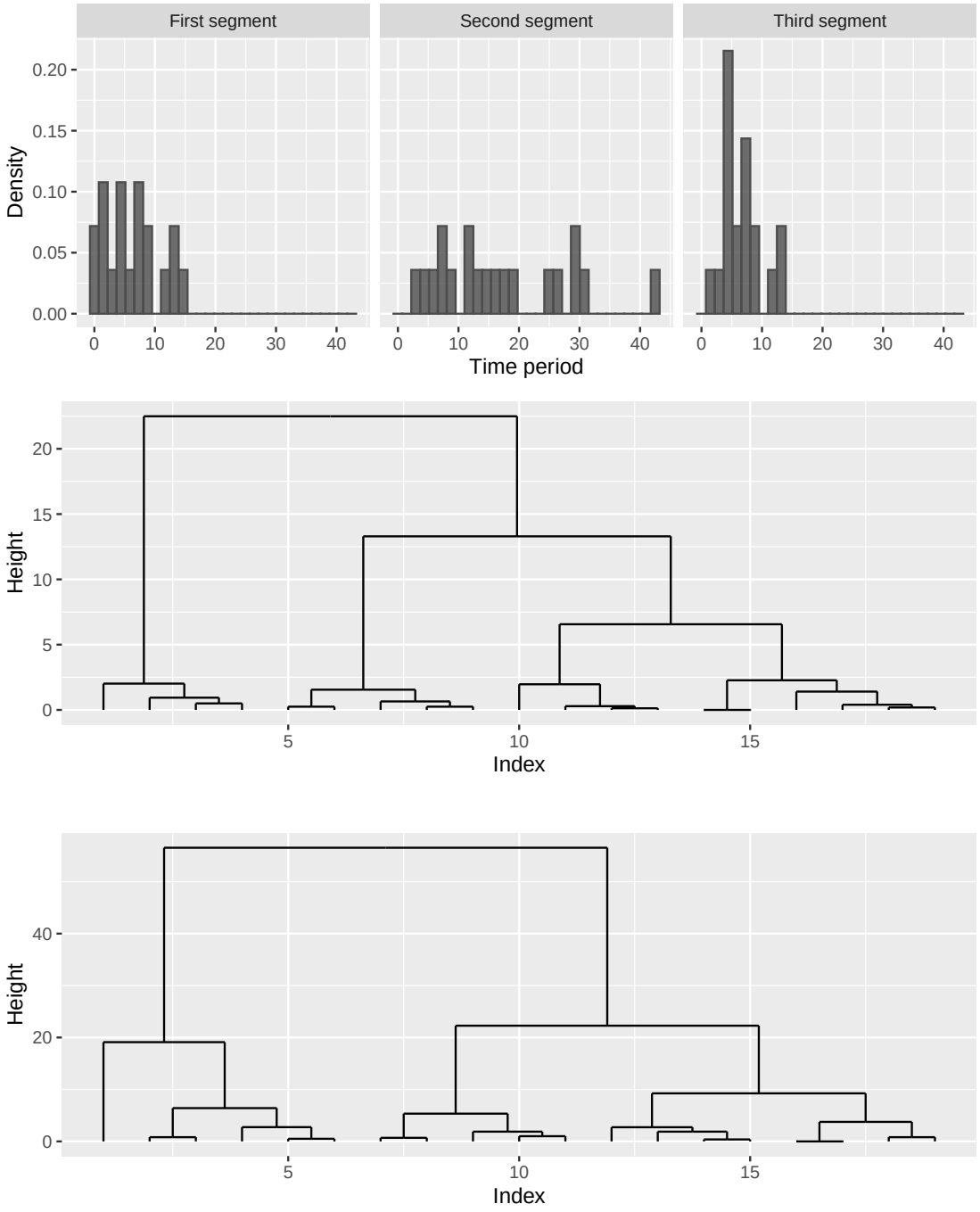


Figure 6.6 – *Analysis of the application of the segmentation/clustering to the Random-Constant context of Experiment I with 2 change-point and 3 clusters. On the top, the time period spent by participants on each segment. On the middle, the dendrogram of the time period spent by participants on the first segment. On the bottom, the dendrogram of the time period spent by participants on the second segment. The data only included participants that moved through the three learning regimes in an ascending order (19 over 22 participants).*

This chapter aims to investigate two methods that allow transfer models to accurate reproduce learning: the segmentation and the segmentation/clustering. Both techniques artificially provide a dynamic to transfer models by partitioning the observations during the learning phase into segments and by estimating the parameters on these segments.

Segmentation Method with Transfer Models

The aim of the segmentation technique is to identify abrupt changes (called change-points) that affect the parameters of the models. The segmentation method was only applied to the GCM on the learning phase of a single experiment (the Random-Constant context of Experiment II). The results showed that all participants improved their performance over time. Moreover, the analysis of the time period spent by participants on the first segment detected two groups of participants with different learning speed (high and low). Finally, the Fisher's exact test of independence performed to determine whether the number of high- and low-speed participants was related to the within-category order was not significant.

Segmentation/Clustering Method with Transfer Models

The segmentation/clustering technique can be considered as a segmentation technique in which the parameters can only take a limited number of values and each segment is associated with a cluster. The segmentation/clustering technique allowed us to *i)* group segments with similar characteristics into a same cluster (facilitating the comparison between participants); and *ii)* increase the number of observations per cluster to better estimate the parameters. The segmentation/clustering method was only applied to the GCM on the learning phase of a single experiment (the Random-Constant context of Experiment II). The segmentation/clustering was applied two times: the first time with 1 change-point and 3 clusters and the second time with 2 change-point and 3 clusters.

Case with 1 change-point and 3 clusters. Three learning regimes were detected: a low/random classification performance, a middle classification performance,

and a high/perfect classification performance. Most of the participants moved from a middle to a high classification performance. Moreover, the majority of participants in the rule-based order moved from a middle to a high classification performance, while the evolution of the participants' performance in the similarity-based order was varied. However, no relation was found between within-category order and evolution of participants' performance. Finally, the analysis of the time period spent by participants on the first segment detected three groups of participants with different learning speed (high, middle, and low).

Case with 2 change-point and 3 clusters. Again, three learning regimes were detected: a low/random classification performance, a middle classification performance, and a high/perfect classification performance. Almost all participants (19 over 22 participants) moved from a low to a middle to a high learning regime. The analysis of the time period spent by participants on the first and second segments detected two groups of participants with different learning speed (high and low). The data of this last analysis only included participants that moved through the three learning regimes in an ascending order.

7

Conclusion and Perspective

Two approaches were developed in the present thesis: an empirical approach and a cognitive modeling approach. The aim of the empirical approach was to investigate the effects that the within-category order (rule-based vs. similarity-based) exerts on learning through a series of laboratory experiments. The cognitive modeling approach aimed at using categorization models to both better understand the mechanism underlying learning and investigate the presentation order.

Empirical Approach *(Chapter 2)*

The analysis of a series of laboratory experiments investigating the within-category order in different contexts showed that *i)* the rule-based order facilitates learning as compared to the similarity-based order when the across-blocks manipulation was constant and categories were either blocked or randomly alternated. *ii)* Participants in the rule-based order showed similar learning performance to participants in the similarity-based order when the across-blocks manipulation was variable and categories were randomly alternated. *iii)* Learning was faster when the across-blocks manipulation was constant and categories were either blocked or randomly alternated (Random-Constant and

Blocked-Constant contexts) as compared to the Random-Variable context, in which the across-blocks manipulation was variable and categories were randomly alternated.

The rule-based order benefited from contexts that encouraged participants to adopt a rule-based strategy. In the Random-Constant context, the constant blocks might have helped participants to focus their attention toward a limited set of information, or might have induced participants to abstract erroneous rules. In both cases, the Random-Constant context promoted the use of a rule-based strategy that enhanced the impact of the rule-based order. In the Blocked-Constant context, the blocked categories amplified the effect of the rule-based order, facilitating the detection of a "principal rule plus exceptions" structure.

Since our dataset did not allow us to conclusively compare different factors (e.g., constant across-blocks manipulation, blocking, etc.), we plan to conduct a full factorial experiment involving the eight following experimental conditions: rule-based vs. similarity-based types $\times$ interleaved vs. blocked categories $\times$ variable vs. constant across-blocks manipulations.

Cognitive Modeling Approach

This part was structured on the duality transfer/learning models. Transfer models are not able to evolve over time and, therefore, they are only adapted to reproduce participants' transfer performance. Conversely, learning models integrate an error-driven mechanism allowing them to reproduce both participants' learning dynamic and transfer performance.

A New Transfer Model *(Chapter 3)*

We developed a new exemplar model based on the Generalized Context Model (GCM) that accounts for the order in which stimuli are presented. This new model, called Ordinal General Context Model(OGCM), was declined into three versions: *i)* the OGCM-L that integrates the average presentation order received during the learning phase, *ii)* the OGCM-M that integrates the most frequent presentation order received during the learning phase, and *iii)* the OGCM-T that integrates the presentation order received during the transfer phase.

Transfer Models Comparison *(Chapter 4)*

The transfer model that best fits the transfer phase of Experiment I was the OGCM-M, an extension of the GCM that accounts for the most frequent stimuli manipulation received during the learning phase. The second model that best fits the experiment data was the OGCM-L, which integrates the average stimuli manipulation received during the learning phase.

The fact that the estimated attention-weight parameter that regulates the ordinal dimension was not negligible in both the OGCM-M and OGCM-L showed that the information provided by the ordinal dimension was relevant for the classification (i.e., the integration of the order received during learning allows models to perform better). Conversely, the estimated ordinal attention-weight parameter was negligible in both the OGCM-T and GCM-Lag, showing that the order received during transfer did not influenced classification. Moreover, the application of the 5-fold cross-validation technique to participants in the rule-based and similarity-based orders (separately) showed that *i)* the models (in particular the OGCM-L and OGCM-M) detected that the majority of the participants in the rule-based order adopted a rule-based strategy, and *ii)* the OGCM-L and OGCM-M were the models that best adapted their predictions to the stimuli manipulation.

However, the goodness-of-fit of the models was similar overall and all models provided good predictions, suggesting that the benefit to integrate the order received during learning is modest (in the studied conditions). We plan to further investigate the role of integrating stimuli manipulation on models by applying an hold-out method in which the parameters are estimated on the learning stimuli and the evaluation is performed on the transfer stimuli.

Learning Models Comparison *(Chapter 5)*

The large size of the dataset of the learning phase allowed us to fit learning models to individual participants' observations. When models are trained on the learning phase and tested on the transfer phase, the Component-Cue model (regardless of its version) best fits the majority of the participants (63-66% Component-Cue vs. 37-34% ALCOVE). All models provided good predictions on most of the items, with the Component-Cue^{L} that provided the best predictions on transfer items and the worse predictions on learning items. When models are trained on the early 80% of the learning phase and tested on the last 20% of the learning phase, ALCOVE (regardless of its version) best fits almost

all participants. Therefore, Component-Cue better captured the generalization patterns (compared to ALCOVE), while ALCOVE better captured the last portion of the learning dynamic (compared to Component-Cue).

Moreover, when models are trained on the learning phase and tested on the transfer phase, Component-Cue best fits a higher number of participants in the rule-based order as compared to participants in the similarity-based order. Conversely, ALCOVE best fits a higher number of participants in the similarity-based order as compared to participants in the rule-based order. However, the difference was not significant. Nonetheless, we found a relation between the type of the model (i.e., Component-Cue vs. ALCOVE) and the within-category order on the generalization patterns. In more detail, we found that *i)* the learning models that best fit participants' performance captured the difference in generalization patterns between participants in the rule-based order and participants in the similarity-based order, and *ii)* the generalization patterns of Component-Cue (when it was the model with the lowest evaluation) were consistent with a rule-based retrieval (confirming that Component-Cue performs better on participants adopting a rule-based strategy).

Finally, we showed that Component-Cue was more sensitive to the stimuli manipulation than ALCOVE. In particular, both versions of Component-Cue showed both distinct learning curves and distinct generalization patterns (at the end of the learning phase) for different stimuli manipulations. Conversely, both versions of ALCOVE showed distinct learning curves for different stimuli manipulations only in the early stage of learning. To further investigate learning models and order manipulation, we plan to conduct classification tasks in which supervised blocks in which order is manipulated are alternated with random unsupervised blocks.

Applying Transfer Models to Learning Data *(Chapter 6)*

Transfer models are only adapted to reproduce transfer performance. Nevertheless, there are a few statistical methods that could allow transfer models to reproduce the learning dynamic. We investigated two of them: the segmentation and the segmentation/clustering. The application of the segmentation method to a single experiment showed that there were two groups of participants: high-speed and low-speed participants. This fact was also confirmed by the application of the segmentation/clustering method to the same experiment.

Moreover, the segmentation/clustering method allowed us to detect a relation between participants' speed and within-category order at the beginning of the classification task. In more detail, participants that showed a high-speed learning in the early stage of the classification task were mostly in the rule-based order. Conversely, participants that showed a low-speed learning in the early stage of the classification task were mostly in the similarity-based order. This shows that the rule-based order is particularly beneficial in the early stage of the classification task. Since time constraints allowed us to apply both techniques to only one experiment, we plan to apply them to all available datasets.

Further Contributions

The use of less known statistical tools as well as the development of a robust inference method represent additional contributions provided by the present thesis.

Statistical Tools

It is common practice to remove from the analysis participants who did not meet the objective of the task. However, the information provided by participants who did not successfully complete the task is relevant. In the present thesis, we promoted the use of two survival analysis techniques that allow researchers to take into account unsuccessful participants: the Kaplan-Meier survival curves and the Cox model. Moreover, all of the statistical tools that we employed are not based on a Gaussian assumption.

Inference Method

A first contribution is represented by the visualization of the set of predictions of the categorization models using the PCA. This practice allows researchers to gain insight into the relations between models. A second contribution is represented by the analysis on the estimation of both the parameters and the classification probability. Having an understanding of the accuracy of the estimation (either parameter estimation or classification probability estimation) allows researchers to better evaluate their results. Finally, the way we applied the hold-out method to participants' learning performance represents the last contribution of the present thesis. Indeed, in machine learning, the hold-out method is applied as follows: after a period of training, models are evaluated by

quantifying the difference between their predictions and feedback. Conversely, in our case, models were evaluated by quantifying the difference between their predictions and participants' responses. In other words, in machine learning feedback serve as training and testing tool, while in our case, feedback (i.e., the effective category) serve as training tool and participants' responses serve as testing tool. The last practice seems to give good results on experimental basis, nevertheless a more rigorous investigation of this new inference method is needed. Additionally, the proposed inference method is being applied to data measuring the brain activity in mice.

We provide the following recommendations to those who wish to fit categorization models to data: *i)* to conduct a preliminary visualization of the selected models using the PCA, *ii)* to study the accuracy of the estimation of either the parameters, or the classification probability, or both, *iii)* to prefer cross-validation method to probabilistic statistical criteria such as BIC or AIC, and *iv)* to ensure through computer simulations that the models are identifiable via the selected cross-validation method.

Bibliography

[Ald97] J. Aldrich. “R. A. Fisher and the making of maximum likelihood 1912-1922”. In: *Statistical Science* (1997) (cit. on p. 155).

[AJT17] D. Altschul, G. Jensen, and H. Terrace. “Perceptual category learning of photographic and painterly stimuli in rhesus macaques (Macaca mulatta) and humans”. In: *PLOS ONE* 12 (Sept. 2017) (cit. on p. 6).

[And91] J. Anderson. “The Adaptive Nature Of Human Categorization”. In: *Psychological Review* 98 (July 1991), pp. 409–429 (cit. on p. 27).

[Ash+98] F. Ashby, L. Alfonso-Reese, A. Turken, and E. Waldron. “A Neuropsychological Theory of Multiple Systems in Category Learning”. In: *Psychological review* 105 (Aug. 1998), pp. 442–81 (cit. on p. 27).

[AM98] F. Ashby and W. Maddox. “Chapter 4. Stimulus Categorization”. In: *Measurement, Judgment, and Decision Making* (Dec. 1998) (cit. on p. 9).

[AM93] F. Ashby and W. Maddox. “Relations between Prototype, Exemplar, and Decision Bound Models of Categorization”. In: *Journal of Mathematical Psychology* 37 (Sept. 1993), pp. 372–400 (cit. on p. 100).

[AL89] I. Auger and C. Lawrence. “Algorithms for the Optimal Identification of Segment Neighborhoods”. In: *Bulletin of mathematical biology* (1989) (cit. on p. 232).

[Bar82] L. Barsalou. “Context-Independent and Context-Dependent Information in Concepts”. In: *Memory & cognition* 10 (Feb. 1982), pp. 82–93 (cit. on p. 6).

[BD62] R. Bellman and S. Dreyfus. *Applied Dynamic Programming*. Princeton University Press, 1962 (cit. on p. 232).

[BH95] Y. Benjamini and Y. Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1995) (cit. on pp. 59, 61).

[Bir+12] M. Birnbaum, N. Kornell, E. Bjork, and R. Bjork. “Why interleaving enhances inductive learning: The roles of discrimination and retrieval”. In: *Memory & cognition* 41 (Nov. 2012) (cit. on pp. 9, 12, 16, 17).

[BS81] K. Bloom and T. Shuell. “Effects of Massed and Distributed Practice on the Learning and Retention of Second-Language Vocabulary”. In: *The Journal of Educational Research* 74 (Mar. 1981), pp. 245–248 (cit. on pp. 15, 16).

[Bou93] M. Bouton. “Context, time, and memory retrieval in the interference paradigms of Pavlovian learning”. In: *Psychological bulletin* 114 (Aug. 1993), pp. 80–99 (cit. on p. 18).

[Bow+69] G. Bower, M. Clark, A. Lesgold, and D. Winzenz. “Hierarchical Retrieval Schemes in Recall of Categorized Word Lists”. In: *Journal of Verbal Learning and Verbal Behavior* 8 (June 1969), pp. 323–343 (cit. on pp. 14, 19).

[Bra08] F. Brady. “The contextual interference effect and sport skills”. In: *Perceptual and motor skills* 106 (May 2008), pp. 461–72 (cit. on p. 15).

[Bro78] L. Brooks. “Nonanalytic concept formation and memory for instances”. In: (Jan. 1978) (cit. on p. 4).

[BGA58] J. Bruner, J. Goodnow, and G. Austin. “A Study of Thinking”. In: *AIBS Bulletin* 7 (Sept. 1958) (cit. on pp. 6, 10).

[BDM84] J. Busemeyer, G. Dewey, and D. Medin. “Evaluation of exemplar-based generalization and the abstraction of categorical information”. In: *Journal of experimental psychology. Learning, memory, and cognition* 10 (Nov. 1984), pp. 638–48 (cit. on p. 28).

[BD10] J. Busemeyer and A. Diederich. *Cognitive Modeling*. Jan. 2010 (cit. on pp. 22–24).

[Car85] S. Carey. “Conceptual Change in Childhood”. In: *Cambridge, MA: Bradford Books* (Jan. 1985) (cit. on p. 6).

[Car+16] K. Carpenter, A. Wills, A. Benattayallah, and F. Milton. “A Comparison of the neural correlates that underlie rule-based and information-integration category learning: COVIS and Category Learning”. In: *Human Brain Mapping* 37 (May 2016) (cit. on p. 10).

[CD05] S. Carpenter and E. DeLosh. “Application of the testing and spacing effects to name learning”. In: *Applied Cognitive Psychology* 19 (July 2005), pp. 619–636 (cit. on p. 16).

[CM13] S. Carpenter and F. Mueller. “The effects of interleaving versus blocking on foreign language pronunciation learning”. In: *Memory & cognition* 41 (Jan. 2013) (cit. on p. 16).

[CG14a] P. F. Carvalho and R. L. Goldstone. “Effects of interleaved and blocked study on delayed test of category learning generalization”. In: *Frontiers in Psychology* 5 (2014), p. 936 (cit. on pp. 9, 12, 17, 30).

[CG14b] P. F. Carvalho and R. L. Goldstone. “Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study”. In: *Memory & Cognition* (2014) (cit. on pp. 9, 11, 12, 16, 17, 30, 35, 36).

[CB12] P. Carvalho and P. B. Albuquerque. “Memory encoding of stimulus features in human perceptual learning”. In: *Journal of Cognitive Psychology* 24 (Oct. 2012), pp. 654–664 (cit. on p. 16).

[CG19] P. Carvalho and R. Goldstone. “A computational model of context-dependent encodings during category learning”. In: (Sept. 2019) (cit. on pp. 21, 31, 139, 165).

[CG11] P. Carvalho and R. Goldstone. “Sequential similarity and comparison effects in category learning”. In: July 2011 (cit. on p. 16).

[CG17] P. Carvalho and R. Goldstone. “The Sequence of Study Changes What Information Is Attended to, Encoded, and Remembered During Category Learning”. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 43 (Mar. 2017) (cit. on pp. 12, 30).

[CG15] P. Carvalho and R. Goldstone. “What you learn is more than what you see: What can sequencing effects tell us about inductive category learning?” In: *Frontiers in Psychology* 6 (Apr. 2015) (cit. on pp. 12, 17).

[CG14c] Paulo Carvalho and Robert Goldstone. “The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study”. In: *Psychonomic bulletin & review* 22 (July 2014) (cit. on pp. 12, 16, 17, 30).

[Cep+09] N. J. Cepeda, N. Coburn, D. Rohrer, et al. “Optimizing Distributed Practice: Theoretical Analysis and Practical Implications”. In: *Experimental psychology* 56 (Feb. 2009), pp. 236–46 (cit. on p. 16).

[CS90] D. L. Cheney and R. M. Seyfarth. *How Monkeys See the World*. University of Chicago Press, 1990 (cit. on p. 3).

[CMB93] S. Choi, M. Mcdaniel, and J. Busemeyer. “Incorporating prior biases in network models of conceptual rule learning”. In: *Memory & cognition* 21 (July 1993), pp. 413–23 (cit. on p. 181).

[Cla14] J. Clapper. “The impact of training sequence and between-category similarity on unsupervised induction”. In: *Quarterly journal of experimental psychology (2006)* 68 (Nov. 2014), pp. 1–55 (cit. on p. 15).

[CN03] A. L. Cohen and R. M. Nosofsky. “An extension of the exemplar-based random-walk model to separable-dimension stimuli”. In: *Journal of Mathematical Psychology* 47.2 (2003), pp. 150–165 (cit. on p. 37).

[Cor+11] K. Corcoran, K. Epstude, L. Damisch, and T. Mussweiler. “Fast Similarities: Efficiency Advantages of Similarity-Focused Comparisons”. In: *Journal of experimental psychology. Learning, memory, and cognition* 37 (June 2011), pp. 1280–6 (cit. on p. 15).

[Cox72] D. R. Cox. “Regression Models and Life-Tables”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1972) (cit. on pp. 55, 56).

[CL72] F. Craik and R. Lockhart. “Levels of Processing: A Framework for Memory Research”. In: *Journal of Verbal Learning and Verbal Behavior* 11 (Dec. 1972), pp. 671– (cit. on p. 23).

[DVS10] P. Delaney, P. Verkoeijen, and A. Spirgel. “Chapter 3 - Spacing and Testing Effects: A Deeply Critical, Lengthy, and At Times Discursive Review of the Literature”. In: *Psychology of Learning and Motivation* 53 (Jan. 2010), pp. 63–147 (cit. on p. 16).

[DG97] S. Dopkins and T. Gleason. “Comparing exemplar and prototype models of categorization”. In: *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 51 (Sept. 1997), pp. 212–230 (cit. on p. 100).

[Dre90] S. E. Dreyfus. “Artificial neural networks, back propagation, and the Kelley-Bryson gradient procedure”. In: *Journal of Guidance Control and Dynamics - J GUID CONTROL DYNAM* 13 (Sept. 1990), pp. 926–928 (cit. on pp. 27, 121).

[Dun+13] J. Dunlosky, K. Rawson, E. Marsh, M. Nathan, and D. Willingham. “Improving Students’ Learning With Effective Learning Techniques”. In: *Psychological Science in the Public Interest* 14 (2013), pp. 4–58 (cit. on p. 29).

[Ebb13] H. Ebbinghaus. “Memory: A Contribution to Experimental Psychology”. In: *NewYork, NY: Teachers College, Columbia University* (1913) (cit. on p. 16).

[EA81] R. Elio and J. Anderson. “The effects of category generalizations and instance similarity on schema abstraction”. In: *Journal of Experimental Psychology: Human Learning and Memory* 7 (Nov. 1981), pp. 397–417 (cit. on pp. 14, 20, 35, 40).

[EA84] R. Elio and J. R. Anderson. “The effects of information order and learning mode on schema abstraction”. In: *Memory & Cognition* (1984) (cit. on pp. 14, 15, 20, 30, 31, 35, 40).

[EK98] M. Erickson and J. Kruschke. “Rules and Exemplars in Category Learning”. In: *Journal of experimental psychology. General* 127 (July 1998), pp. 107–40 (cit. on p. 27).

[Est86] W. Estes. “Array Models for Category Learning”. In: *Cognitive psychology* 18 (Nov. 1986), pp. 500–49 (cit. on p. 26).

[Est94] W. K. Estes. *Classification and Cognition*. New York, NY: Oxford University Press, 1994 (cit. on p. 4).

[FTT15] J. Fan, N. Turk-Browne, and J. Taylor. “Error-Driven Learning in Statistical Summary Perception”. In: *Journal of experimental psychology. Human perception and performance* 42 (Sept. 2015) (cit. on p. 29).

[FL10] S. Farrell and S. Lewandowsky. “Computational Models as Aids to Better Reasoning in Psychology”. In: *Current Directions in Psychological Science - CURR DIRECTIONS PSYCHOL SCI* 19 (Oct. 2010), pp. 329–335 (cit. on p. 24).

[Fel03] J. Feldman. “A Catalog of Boolean Concepts”. In: *J. Math. Psychol.* (2003) (cit. on pp. 68, 69, 74).

[Fel00] J. Feldman. “Minimization of Boolean complexity in human concept learning”. In: *Nature* 407 (Nov. 2000), pp. 630–3 (cit. on p. 27).

[Fox+00] E. Fox, V. Lester, R. Russo, et al. “Facial Expressions of Emotion: Are Angry Faces Detected More Efficiently?” In: *Cognition & Emotion* (2000) (cit. on p. 3).

[Gag50] R. Gagné. “The effect of sequence of presentation of similar items on the learning of paired associates”. In: *Journal of Experimental Psychology* 40 (Feb. 1950), pp. 61–73 (cit. on pp. 9, 12).

[Gar53] W. R. Garner. “An informational analysis of absolute judgments of loudness”. In: *Journal of Experimental Psychology* 46 (1953), pp. 373–380 (cit. on p. 18).

[Gar74] W. R. Garner. *The processing of information and structure.* 1974 (cit. on p. 104).

[GBD92] S. Geman, E. Bienenstock, and R. Doursat. “Neural Networks and the Bias/Variance Dilemma”. In: *Neural Computation* 4 (Jan. 1992), pp. 1–58 (cit. on p. 143).

[Gle76] A. Glenberg. “Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms”. In: *Journal of Verbal Learning and Verbal Behavior* 15 (Feb. 1976), pp. 1–16 (cit. on p. 16).

[GL80] A. Glenberg and T. Lehmann. “Spacing repetitions over 1 week”. In: *Memory & cognition* 8 (Dec. 1980), pp. 528–38 (cit. on p. 16).

[GBH89] M. A. Gluck, G. Bower, and M. R. Hee. “A configural-cue network model of animal and human associative learning”. In: *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (Jan. 1989), pp. 323–332 (cit. on pp. 27, 121, 122, 128, 181).

[GB88a] M. Gluck and G. Bower. “Evaluating an adaptive network model of human learning”. In: *Journal of Memory and Language - J MEM LANG* 27 (Apr. 1988), pp. 166–195 (cit. on pp. 121, 130, 181).

[GB88b] M. Gluck and G. Bower. “From Conditioning to Category Learning: An Adaptive Network Model”. In: *Journal of experimental psychology. General* (1988) (cit. on pp. 21, 23, 25, 27, 32, 33, 121, 122, 125, 130, 181).

[Gol94] R. Goldstone. “The role of similarity in categorization: Providing a groundwork”. In: *Cognition* 52 (Sept. 1994), pp. 125–57 (cit. on p. 6).

[Gol96] R. L. Goldstone. “Isolated and interrelated concepts”. In: *Memory & Cognition* (1996) (cit. on pp. 9, 16, 35).

[GKC12] R. Goldstone, A. Kersten, and P. Carvalho. “Concepts and categorization”. In: Oct. 2012, pp. 607–630 (cit. on p. 4).

[GLS01] R. Goldstone, Y. Lippa, and R. Shiffrin. “Altering object representations through category learning”. In: *Cognition* 78 (Feb. 2001), pp. 27–43 (cit. on p. 9).

[Goo72] N. Goodman. “Seven Strictures on Similarity”. In: *Problems and Projects* (Jan. 1972) (cit. on p. 6).

[GW06] D. Goren and H. Wilson. “Quantifying facial expression recognition across viewing conditions”. In: *Vision research* (2006) (cit. on p. 3).

[GV60] W. Gorman and S. Valavanis. “Econometrics: An Introduction to Maximum Likelihood Methods.” In: *Journal of the Royal Statistical Society. Series A (General)* 123 (Jan. 1960), p. 494 (cit. on p. 24).

[HGV11] A. Helsdingen, T. Gog, and J. J. G. Van Merrienboer. “The effects of practice schedule on learning a complex judgment task”. In: *Learning and Instruction* 21 (Feb. 2011), pp. 126–136 (cit. on p. 15).

[HH52] A.L. Hodgkin and A.F. Huxley. “A Quantitative Description Of Membrane Current And Its Application To Conduction And Excitation In Nerve”. In: *The Journal of physiology* 117 (Sept. 1952), pp. 500–44 (cit. on p. 23).

[HL68] M. Holland and G. Lockhead. “Sequential effects in absolute judgments of loudness”. In: *Perception & Psychophysics* 3 (Nov. 1968), pp. 409–414 (cit. on p. 18).

[HK01] M. Howard and M. Kahana. “A Distributed Representation of Temporal Context”. In: *Journal of Mathematical Psychology* 46 (Nov. 2001), pp. 269–299 (cit. on p. 21).

[Hsu+19] A. Hsu, J. Martin, A. Sanborn, and T. Griffiths. “Identifying category representations for complex stimuli using discrete Markov chain Monte Carlo with people”. In: *Behavior Research Methods* 51 (Feb. 2019) (cit. on p. 21).

[HY12] S.-M. Hsu and L.-X. Yang. “Sequential Effects in Facial Expression Categorization”. In: *Emotion (Washington, D.C.)* 13 (Apr. 2012) (cit. on p. 18).

[JOU11] M. Jitsumori, M. Ohkita, and T. Ushitani. “The learning of basic-level categories by pigeons: The prototype effect, attention, and effects of categorization”. In: *Learning & behavior* 39 (Apr. 2011), pp. 271–87 (cit. on p. 6).

[JK05] M. Johansen and J. Kruschke. “Category Representation for Classification and Feature Inference.” In: *Journal of experimental psychology. Learning, memory, and cognition* 31 (Dec. 2005), pp. 1433–58 (cit. on p. 37).

[JP03] M. Johansen and T. Palmeri. “Are there representational shifts in category learning?” In: *Cognitive psychology* 45 (Jan. 2003), pp. 482–553 (cit. on p. 37).

[JLM06] M. Jones, B. Love, and W. Maddox. “Recency effects as a window to generalization: Separating decisional and perceptual sequential effects in category learning”. In: *Journal of experimental psychology. Learning, memory, and cognition* 32 (Apr. 2006), pp. 316–32 (cit. on pp. 18, 19).

[JS03] M. Jones and W. Sieck. “Learning Myopia: An Adaptive Recency Effect in Category Learning”. In: *Journal of experimental psychology. Learning, memory, and cognition* 29 (Aug. 2003), pp. 626–40 (cit. on pp. 15, 18).

[KP12] Sean H. K. Kang and Harold Pashler. “Learning Painting Styles: Spacing is Advantageous when it Promotes Discriminative Contrast”. In: 2012 (cit. on pp. 9, 12, 16, 17, 35, 36).

[KM58] E. L. Kaplan and P. Meier. “Nonparametric Estimation from Incomplete Observations”. In: *Journal of the American Statistical Association* (1958) (cit. on pp. 52, 54).

[KB11] J. Karpicke and A. Bauernschmidt. “Spaced Retrieval: Absolute Spacing Enhances Learning Regardless of Relative Spacing”. In: *Journal of experimental psychology. Learning, memory, and cognition* 37 (May 2011), pp. 1250–7 (cit. on p. 16).

[KR10] J. Karpicke and H. Roediger. “Is expanding retrieval a superior method for learning text materials?” In: *Memory & cognition* 38 (Jan. 2010), pp. 116–24 (cit. on p. 16).

[KW97] R. Kohavi and D. Wolpert. “Bias Plus Variance Decomposition for Zero-One Loss Functions”. In: (Sept. 1997) (cit. on p. 143).

[Kom92] L. Komatsu. “Recent Views Of Conceptual Structure”. In: *Psychological Bulletin* 112 (Nov. 1992), pp. 500–526 (cit. on p. 6).

[KB08] N. Kornell and R. Bjork. “Learning concepts and categories: is spacing the "enemy of induction"?” In: *Psychological science* (2008) (cit. on pp. 9, 12, 16, 17, 35, 36).

[Kor+10] N. Kornell, A. Castel, T. Eich, and R. Bjork. “Spacing as the friend of both memory and induction in young and older adults”. In: *Psychology and aging* 25 (June 2010), pp. 498–503 (cit. on pp. 12, 16).

[KV18] N. Kornell and K. Vaughn. “In Inductive Category Learning, People Simultaneously Block and Space Their Studying Using a Strategy of Being Thorough and Fair”. In: *Archives of Scientific Psychology* 6 (Nov. 2018), pp. 138–147 (cit. on pp. 9, 12).

[KCG15] A. S. Kost, P. F. Carvalho, and R. L. Goldstone. “Can You Repeat That ? The Effect of Item Repetition on Interleaved and Blocked Study”. In: *Proceedings of the 37th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society* (2015), pp. 1189–1194 (cit. on pp. 16, 17).

[Kru01] J. Kruschke. “Toward a Unified Model of Attention in Associative Learning”. In: *Journal of Mathematical Psychology* 45 (Dec. 2001), pp. 812–863 (cit. on p. 26).

[Kru92] J. K. Kruschke. “ALCOVE: an exemplar-based connectionist model of category learning.” In: *Psychological review* (1992) (cit. on pp. 21, 23, 26–28, 32, 33, 121, 131, 134, 181).

[Kru05] J. K. Kruschke. “Category learning”. In: *The Handbook of Cognition* (Jan. 2005) (cit. on p. 4).

[Kru08] J. K. Kruschke. “Models of categorization”. In: *The Cambridge handbook of computational psychology* (Jan. 2008), pp. 267–301 (cit. on p. 25).

[KJ99] J. Kruschke and M. Johansen. “A Model of Probabilistic Category Learning”. In: *Journal of experimental psychology. Learning, memory, and cognition* 25 (Oct. 1999), pp. 1083–119 (cit. on p. 27).

[KW78] J. Kruskal and M. Wish. “Multidimensional Scaling”. In: *Thousand Oaks, CA: Sage* 11 (1978) (cit. on p. 100).

[KH56] K. Kurtz and C. Hovland. “Concept Learning with Different Sequences of Instances”. In: *Journal of experimental psychology* 51 (May 1956), pp. 239–43 (cit. on p. 16).

[Lac97] Y. Lacouture. “Bow, range, and sequential effects in absolute identification: A response-time analysis”. In: *Psychological research* 60 (Feb. 1997), pp. 121–33 (cit. on p. 18).

[LLM07] D. Lafond, Y. Lacouture, and G. Mineau. “Complexity minimization in rule-based category learning: Revising the catalog of Boolean concepts and evidence for non-minimal rules”. In: *Journal of Mathematical Psychology* 51 (Apr. 2007), pp. 57–74 (cit. on p. 37).

[Lam00] K. Lamberts. “Information accumulation theory of categorization”. In: *Psychological review* 107 (May 2000), pp. 227–60 (cit. on p. 37).

[Lav05] M. Lavielle. “Using penalized contrasts for the change-point problem”. In: *Signal Processing* (2005) (cit. on p. 232).

[LN02] M. Lee and D. Navarro. “Extending the ALCOVE model of category learning to featural stimulus domains”. In: *Psychonomic bulletin & review* 9 (Apr. 2002), pp. 43–58 (cit. on p. 181).

[Leg05] A. Legendre. *Nouvelles Méthodes Pour la Détermination des Orbites des Comètes*. Jan. 1805 (cit. on p. 24).

[Lev75] M. Levine. “A cognitive theory of learning. Research on hypothesis testing”. In: (Jan. 1975) (cit. on p. 27).

[LCK12] N. Li, W. Cohen, and K. Koedinger. “Problem Order Implications for Learning”. In: *International Journal of Artificial Intelligence in Education* 23 (June 2012) (cit. on pp. 15, 16).

[Lip61] L. Lipsitt. “Simultaneous and Successive Discrimination Learning in Children”. In: *Child development* 32 (July 1961), pp. 337–47 (cit. on p. 15).

[LB08] J. Logan and D. Balota. “Expanded vs. Equal Interval Spaced Retrieval Practice: Exploring Different Schedules of Spacing and Retention Interval in Younger and Older Adults”. In: *Neuropsychology, development, and cognition. Section B, Aging, neuropsychology and cognition* 15 (June 2008), pp. 257–80 (cit. on p. 16).

[LMG04] B. Love, D. Medin, and T. Gureckis. “SUSTAIN: a network model of category learning”. In: *Psychological review* 111 (May 2004), pp. 309–32 (cit. on pp. 21, 27).

[Luc63] R. D. Luce. “Detection and recognition”. In: *Handbook of mathematical psychology*. Ed. by & E. Galanter (Eds.) In R. D. Luce R. R. Bush. New York: Wiley, 1963, pp. 103–189 (cit. on p. 111).

[LS08] U. Luxburg and B. Schoelkopf. “Statistical Learning Theory: Models, Concepts, and Results”. In: *Handbook of the History of Logic* 10 (Nov. 2008) (cit. on p. 143).

[MP15] M. Mack and T. Palmeri. “The Dynamics of Categorization: Unraveling Rapid Categorization”. In: *Journal of experimental psychology. General* 144 (May 2015) (cit. on pp. 15, 29).

[Mac75] N. Mackintosh. “A theory of attention: Variations in the associability of stimuli with reinforcement”. In: *Psychological Review* 82 (July 1975), pp. 276–298 (cit. on p. 26).

[Mad+10] W. Maddox, J. Pacheco, M. Reeves, B. Zhu, and D. Schnyer. “Rule-Based and Information-Integration Category Learning in Normal Aging”. In: *Neuropsychologia* 48 (Aug. 2010), pp. 2998–3008 (cit. on p. 10).

[MF90] D. Massaro and D. Friedman. “Models of integration given multiple sources of information”. In: *Psychological review* 97 (May 1990), pp. 225–52 (cit. on p. 26).

[MF09] F. Mathy and J. Feldman. “A rule-based presentation order facilitates category learning”. In: *Psychonomic Bulletin & Review* (2009) (cit. on pp. 9, 10, 14, 20, 21, 30, 31, 35).

[MF16] F. Mathy and J. Feldman. “The Influence of Presentation Order on Category Transfer.” In: *Experimental psychology* (2016) (cit. on pp. 9, 11, 14, 20, 21, 30, 31, 35, 37, 45, 62, 63, 182, 214).

[MFP13] M. Mcdaniel, C. Fadler, and H. Pashler. “Effects of Spaced Versus Massed Raining in Function Learning”. In: *Journal of experimental psychology. Learning, memory, and cognition* 39 (Apr. 2013) (cit. on p. 15).

[MN95] S. McKinley and R. Nosofsky. “Investigations of Exemplar and Decision Bound Models in Large, Ill-Defined Category Structures”. In: *Journal of experimental psychology. Human perception and performance* 21 (Mar. 1995), pp. 128–48 (cit. on p. 33).

[Mea+17] B. Meagher, P. Carvalho, R. Goldstone, and R. Nosofsky. “Organized simultaneous displays facilitate learning of complex natural science categories”. In: *Psychonomic Bulletin & Review* 24 (Feb. 2017) (cit. on pp. 9, 30).

[Mea+18] B. Meagher, K. Cataldo, B. Douglas, M. Mcdaniel, and R. Nosofsky. “Training of rock classifications: The use of computer images versus physical rock samples”. In: *Journal of Geoscience Education* 66 (May 2018), pp. 1–10 (cit. on p. 9).

[MDM83] D. L. Medin, G. I. Dewey, and T. D. Murphy. “Relationships between item and category learning: Evidence that abstraction is not automatic.” In: *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9(4), 607–625.* (1983) (cit. on p. 104).

[MS78] D. L. Medin and M. M. Schaffer. “Context theory of classification learning.” In: *Psychological Review, 85 (3): 207-238.* (1978) (cit. on pp. 4, 21, 26, 37, 103, 104, 106, 113, 181).

[MB94] D. Medin and J. Bettger. “Presentation order and recognition of categorically related examples”. In: *Psychonomic Bulletin & Review* 1 (June 1994), pp. 250–254 (cit. on pp. 14, 20).

[MC98] D. Medin and J. Coley. “Concepts and categorization”. In: *Perception and Cognition at Century’s End* (Jan. 1998), pp. 403–439 (cit. on p. 4).

[MGG93] D. Medin, R. Goldstone, and D. Gentner. “Respects for Similarity”. In: *Psychological Review* 100 (Apr. 1993), pp. 254–278 (cit. on p. 6).

[Med+82] D.L. Medin, M.W. Altom, S.M. Edelson, and D. Freko. “Correlated symptoms and simulated medical classification.” In: *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8, 37–50.* (1982) (cit. on p. 104).

[MR81] C. B. Mervis and E. Rosch. “Categorization of Natural Objects”. In: *Annual Review of Psychology* 32 (Nov. 1981), pp. 89–115 (cit. on p. 5).

[MS02] J. Minda and J. Smith. “Comparing Prototype-Based and Exemplar-Based Accounts of Category Learning and Attentional Allocation”. In: *Journal of experimental psychology. Learning, memory, and cognition* 28 (Apr. 2002), pp. 275–92 (cit. on p. 37).

[MNH08] C. Mitchell, S. Nash, and G. Hall. “The Intermixed-Blocked Effect in Human Perceptual Learning Is Not the Consequence of Trial Spacing”. In: *Journal of experimental psychology. Learning, memory, and cognition* 34 (Feb. 2008), pp. 237–42 (cit. on p. 17).

[Miy+18] T. Miyatsu, R. Gouravajhala, R. Nosofsky, and M. Mcdaniel. “Feature Highlighting Enhances Learning of a Complex Natural-Science Category”. In: *Journal of experimental psychology. Learning, memory, and cognition* (Apr. 2018) (cit. on p. 9).

[Mor89] S. Mori. “A limited-capacity response process in absolute identification”. In: *Perception & psychophysics* 46 (Sept. 1989), pp. 167–73 (cit. on p. 18).

[Mun+10] P. Munro, H. Toivonen, G. Webb, et al. “Bias Variance Decomposition”. In: Jan. 2010 (cit. on p. 143).

[Mur62] B. B. Murdock. “The serial position effect in free recall”. In: *Journal of Experimental Psychology* 106 (Jan. 1962), pp. 226–254 (cit. on p. 18).

[Mur11] G. L. Murphy. “The contribution (and drawbacks) of models to the study of concepts”. In: *Formal Approaches in Categorization*. Ed. by E. M. Pothos and A. J. Editors Wills. Cambridge University Press, 2011, pp. 299–312 (cit. on pp. 23, 24).

[MM85] G. Murphy and D. Medin. “The Role Theories in Conceptual Coherence”. In: *Psychological review* 92 (Aug. 1985), pp. 289–316 (cit. on p. 6).

[Mye76] J. L. Myers. “Probability learning”. In: *In W. K. Estes (Ed.),Handbook of learning and cognitive processes: Vol. 3. Approaches to human learning and motivation(pp. 171–205). Hillsdale, NJ: Erlbaum* (1976) (cit. on p. 18).

[Noh+16] S. Noh, V. Yan, R. Bjork, and W. Maddox. “Optimal sequencing during category learning: Testing a dual-learning systems perspective”. In: *Cognition* 155 (June 2016), pp. 23–29 (cit. on pp. 9, 10, 12, 17, 36).

[Nor05] D. Norris. “How do computational models help us develop better theories?” In: (Jan. 2005) (cit. on p. 24).

[Nos87] R. M. Nosofsky. “Attention and learning processes in the identification and categorization of integral stimuli.” In: *Journal of experimental psychology. Learning, memory, and cognition.* (1987) (cit. on pp. 4, 10).

[Nos86] R. M. Nosofsky. “Attention, similarity, and the identification–categorization relationship.” In: *Journal of Experimental Psychology: General, 115 (1): 39-57.* (1986) (cit. on pp. 9–11, 21, 25, 26, 28, 33, 99, 100, 103, 113).

[Nos84] R. M. Nosofsky. “Choice, similarity, and the context theory of classification.” In: *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10(1), 104–114.* (1984) (cit. on pp. 4, 103, 111–113).

[Nos85] R. M. Nosofsky. “Overall similarity and the identification of separable-dimension stimuli: A choice model analysis.” In: *Perception & Psychophysics 38 (5): 415-432.* (1985) (cit. on p. 106).

[Nos11] R. M. Nosofsky. “The generalized context model: An exemplar model of classification”. In: *Formal approaches in categorization* (Jan. 2011), pp. 18–39 (cit. on p. 33).

[NP98] R. M. Nosofsky and T. Palmeri. “A rule-plus-exception model for classifying objects in continuous-dimension spaces”. In: *Psychon. Bull. Rev.* 5 (Sept. 1998), pp. 345–369 (cit. on p. 27).

[NPM94] R. M. Nosofsky, T. Palmeri, and S. McKinley. “Rule-plus-exception model of classification learning”. In: *Psychological review* 101 (Feb. 1994), pp. 53–79 (cit. on p. 27).

[NSM17] R. M. Nosofsky, C. A. Sanders, and M. A. McDaniel. “Tests of an Exemplar-Memory Model of Classification Learning in a High-Dimensional Natural-Science Category Domain”. In: *Journal of Experimental Psychology: General* 147 (Oct. 2017) (cit. on pp. 21, 100, 103, 139).

[NSM18a] R. M. Nosofsky, C. Sanders, and M. Mcdaniel. “A Formal Psychological Model of Classification Applied to Natural-Science Category Learning”. In: *Current Directions in Psychological Science* 27 (Mar. 2018), p. 096372141774095 (cit. on pp. 21, 100, 101, 103).

[Nos+18] R. M. Nosofsky, C. Sanders, X. Zhu, and M. Mcdaniel. “Model-guided search for optimal natural-science-category training exemplars: A work in progress”. In: *Psychonomic Bulletin & Review* 26 (July 2018) (cit. on pp. 21, 24, 100, 103, 139, 140, 165).

[Nos+94] R. Nosofsky, M. Gluck, T. Palmeri, S. Mckinley, and P. Glauthier. “Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961)”. In: *Memory & cognition* (1994) (cit. on pp. 121, 134, 139, 165, 181).

[NKM92] R. Nosofsky, J. Kruschke, and S. McKinley. “Combining Exemplar-Based Category Representations and Connectionist Learning Rules”. In: *Journal of experimental psychology. Learning, memory, and cognition* (1992) (cit. on pp. 33, 100, 103, 118, 119, 121, 128, 134, 139, 165, 181).

[NP15] R. Nosofsky and T. Palmeri. “An exemplar-based random-walk model of categorization and recognition”. In: *Oxford Handbook Computational and Mathematical Psychology* (Jan. 2015), pp. 142–164 (cit. on p. 26).

[NP96] R. Nosofsky and T. Palmeri. “Learning to classify integral-dimension stimuli”. In: *Psychonomic bulletin & review* 3 (June 1996), pp. 222–226 (cit. on pp. 9, 10, 101).

[NSM18b] R. Nosofsky, C. Sanders, and M. Mcdaniel. “A Formal Psychological Model of Classification Applied to Natural-Science Category Learning”. In: *Current Directions in Psychological Science* (2018) (cit. on pp. 100, 103).

[Nos+19] R. Nosofsky, C. Sanders, B. Meagher, and B. Douglas. “Search for the Missing Dimensions: Building a Feature-Space Representation for a Natural-Science Category Domain”. In: *Computational Brain & Behavior* (June 2019), pp. 1–21 (cit. on p. 9).

[Nos+17] R. Nosofsky, C. Sanders, B. Meagher, and B. Douglas. “Toward the development of a feature-space representation for a complex natural category domain”. In: *Behavior research methods* 50 (Apr. 2017) (cit. on p. 100).

[NS05] R. Nosofsky and R. Stanton. “Speeded Classification in a Probabilistic Category Structure: Contrasting Exemplar-Retrieval, Decision-Boundary, and Prototype Models.” In: *Journal of experimental psychology. Human perception and performance* 31 (July 2005), pp. 608–29 (cit. on p. 100).

[NZ02] R. Nosofsky and S. Zaki. “Exemplar and Prototype Models Revisited: Response Strategies, Selective Attention, and Stimulus Generalization”. In: *Journal of experimental psychology. Learning, memory, and cognition* 28 (Oct. 2002), pp. 924–40 (cit. on p. 100).

[Pal99] T. Palmeri. “Learning categories at different hierarchical levels: A comparison of category learning models”. In: *Psychonomic bulletin & review* 6 (Sept. 1999), pp. 495–503 (cit. on pp. 121, 134, 165, 181).

[PM15] T. Palmeri and M. Mack. “How Experimental Trial Context Affects Perceptual Categorization”. In: *Frontiers in psychology* 6 (Feb. 2015), p. 180 (cit. on p. 29).

[Pas+07] H. Pashler, D. Rohrer, M. Wiseheart, and S. Carpenter. “Enhancing learning and retarding forgetting: Choices and consequences”. In: *Psychonomic bulletin & review* 14 (May 2007), pp. 187–93 (cit. on p. 16).

[Pic+07] F. Picard, S. Robin, E. Lebarbier, and J.-J. Daudin. “A Segmentation/Clustering Model for the Analysis of Array CGH Data”. In: *Biometrics* 63 (2007), pp. 758–66 (cit. on pp. 238–240).

[PM02] M. Pitt and I. Myung. “When a good fit can be bad”. In: *Trends in cognitive sciences* 6 (Nov. 2002), pp. 421–425 (cit. on p. 145).

[Pop59] K. Popper. *Logic of Scientific Discovery*. Jan. 1959 (cit. on p. 22).

[PC02] E. Pothos and N. Chater. “A Simplicity Principle in unsupervised human categorization”. In: *Cognitive Science* 26 (May 2002), pp. 303–343 (cit. on p. 27).

[PW11] E. Pothos and A. Wills. *Formal Approaches in Categorization*. Jan. 2011 (cit. on p. 4).

[PFM09] A. Price, J. Filoteo, and W. Maddox. “Rule Based Category Learning in Patients with Parkinson’s Disease”. In: *Neuropsychologia* 47 (May 2009), pp. 1213–26 (cit. on p. 10).

[QA14] T. Qian and R. Aslin. “Learning bundles of stimuli renders stimulus order as a cue, not a confound”. In: *Proceedings of the National Academy of Sciences of the United States of America* 111 (Sept. 2014) (cit. on p. 15).

[RTJ14] K. Rawson, R. Thomas, and L. Jacoby. “The Power of Examples: Illustrative Examples Enhance Conceptual Learning of Declarative Concepts”. In: *Educational Psychology Review* 27 (June 2014) (cit. on pp. 16, 17).

[Ree72] S. Reed. “Pattern Recognition and Categorization”. In: *Cognitive Psychology* 3 (July 1972), pp. 382–407 (cit. on p. 26).

[RH05] B. Rehder and A. Hoffman. “Thirty-something categorization results explained: Attention, eyetracking, and models of category learning”. In: *Journal of experimental psychology. Learning, memory, and cognition* 31 (Oct. 2005), pp. 811–29 (cit. on pp. 21, 37, 100, 103).

[RLP08] T. Rickard, J. Lau, and H. Pashler. “Spacing and the transition from calculation to retrieval”. In: *Psychonomic bulletin & review* 15 (July 2008), pp. 656–61 (cit. on p. 16).

[Rip89] L. Rips. “Similarity, typicality, and categorization”. In: *Similarity and analogical reasoning* (Nov. 1989), pp. 21–59 (cit. on p. 6).

[Roh12] D. Rohrer. “Interleaving Helps Students Distinguish among Similar Concepts”. In: *Educational Psychology Review* 24 (Sept. 2012), pp. 355–367 (cit. on p. 16).

[Roh09] D. Rohrer. “The Effects of Spacing and Mixing Practice Problems”. In: *Journal for Research in Mathematics Education* 40 (Jan. 2009), pp. 4–17 (cit. on p. 16).

[Ros75] E. Rosch. “Cognitive Representation of Semantic Categories”. In: *Journal of Experimental Psychology: General* 104 (Sept. 1975), pp. 192–233 (cit. on p. 5).

[RM75] E. Rosch and C. Mervis. “Family Resemblances: Studies in the Internal Structure of Categories”. In: *Cognitive Psychology* 7 (Oct. 1975), pp. 573–605 (cit. on pp. 5, 26).

[Ros58] F. Rosenblatt. “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain”. In: *Psychological Review* 65 (Jan. 1958), pp. 386– (cit. on pp. 27, 121).

[RR04] J. Rouder and R. Ratcliff. “Comparing Categorization Models”. In: *Journal of experimental psychology. General* 133 (Apr. 2004), pp. 63–82 (cit. on pp. 21, 103, 181).

[SJL08] Y. Sakamoto, M. Jones, and B. Love. “Putting the psychology back into psychological models: Mechanistic versus rational approaches”. In: *Memory & cognition* 36 (Oct. 2008), pp. 1057–65 (cit. on p. 21).

[Sam69] S. Samuels. “Effect of simultaneous versus successive discrimination training on paired-associate learning”. In: *Journal of Educational Psychology* 60 (Feb. 1969), pp. 46–48 (cit. on p. 15).

[SYK16] F. Sana, V. Yan, and J. Kim. “Study Sequence Matters for the Inductive Learning of Cognitive Concepts”. In: *Journal of Educational Psychology* 109 (Apr. 2016) (cit. on p. 16).

[SN20] C. Sanders and R. Nosofsky. “Training Deep Networks to Construct a Psychological Feature Space for a Natural-Object Category Domain”. In: *Computational Brain & Behavior* (Jan. 2020), pp. 1–23 (cit. on pp. 21, 100, 103, 139).

[SD08] C. Sandhofer and L. Doumas. “Order of Presentation Effects in Learning Color Categories”. In: *Journal of Cognition and Development - J COGN DEV* 9 (Apr. 2008), pp. 194–221 (cit. on p. 15).

[SP15] M. Schlichting and A. Preston. “Memory integration: Neural mechanisms and implications for behavior”. In: *Current Opinion in Behavioral Sciences* 1 (Feb. 2015), pp. 1–8 (cit. on p. 29).

[She80] R. Shepard. “Multidimensional Scaling, Tree-Fitting, and Clustering”. In: *Science (New York, N.Y.)* 210 (Nov. 1980), pp. 390–8 (cit. on p. 100).

[She57] R. Shepard. “Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space”. In: *Psychometrika* 22 (Dec. 1957), pp. 325–345 (cit. on p. 111).

[She64] R. N. Shepard. “Attention and the metric structure of the stimulus space.” In: *Journal of Mathematical Psychology, 1 (1): 54-87.* (1964) (cit. on pp. 104, 112).

[She87] R. N. Shepard. “Toward a universal law of generalization for psychological science.” In: *Science. 237 (4820): 1317-1323.* (1987) (cit. on p. 106).

[SHJ61] R. N. Shepard, C. I. Hovland, and H. M. Jenkins. “Learning and memorization of classifications.” In: *Psychological Monographs: General and Applied, 75(13), 1–42.* (1961) (cit. on pp. 9, 10, 27, 101, 112).

[SS97] R. Shiffrin and M. Steyvers. “A model for recognition memory: REM - Retrieving effectively from memory”. In: *Psychonomic bulletin & review* 4 (June 1997), pp. 145–66 (cit. on p. 23).

[Slo96] S. Sloman. “The Empirical Case For Two Systems of Reasoning”. In: *Psychological Bulletin* 119 (Jan. 1996), pp. 3– (cit. on p. 20).

[SSO93] E. Smith, E. Shafir, and D. Osherson. “Similarity, plausibility, and judgments of probability”. In: *Cognition* 49 (Oct. 1993), pp. 67–96 (cit. on p. 6).

[SCR10] J. Smith, W. Chapman, and J. Redford. “Stages of Category Learning in Monkeys (Macaca mulatta) and Humans (Homo sapiens)”. In: *Journal of experimental psychology. Animal behavior processes* 36 (Jan. 2010), pp. 39–53 (cit. on p. 9).

[SE15] J. Smith and S. Ell. “One Giant Leap for Categorizers: One Small Step for Categorization Theory”. In: *PloS one* 10 (Sept. 2015), e0137334 (cit. on p. 181).

[SM98] J. Smith and J. Minda. “Prototypes in the mist: The early epochs of category learning”. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24 (Nov. 1998), pp. 1411–1436 (cit. on p. 100).

[SM00] J. Smith and J. Minda. “Thirty Categorization Results in Search of a Model”. In: *Journal of experimental psychology. Learning, memory, and cognition* 26 (Feb. 2000), pp. 3–27 (cit. on pp. 21, 28, 37, 103).

[Smi+16] J. Smith, A. Zakrzewski, J. Johnson, J. Valleau, and B. Church. “Categorization: The View from Animal Cognition”. In: *Behavioral Sciences* 6 (June 2016), p. 12 (cit. on p. 6).

[SA08] B. Spiering and F. Ashby. “Response Processes in Information-Integration Category Learning”. In: *Neurobiology of learning and memory* 90 (July 2008), pp. 330–8 (cit. on p. 10).

[SB04] N. Stewart and G. Brown. “Sequence Effects in the Categorization of Tones Varying in Frequency”. In: *Journal of experimental psychology. Learning, memory, and cognition* 30 (Apr. 2004), pp. 416–30 (cit. on p. 18).

[SBC02] N. Stewart, G. Brown, and N. Chater. “Sequence Effects in Categorization of Simple Perceptual Stimuli”. In: *Journal of experimental psychology. Learning, memory, and cognition* 28 (Feb. 2002), pp. 3–11 (cit. on pp. 18, 19).

[SDR00] G. Storms, P. De Boeck, and W. Ruts. “Prototype and Exemplar-Based Information in Natural Language Categories”. In: *Journal of Memory and Language - J MEM LANG* 42 (Jan. 2000), pp. 51–73 (cit. on p. 100).

[SOS92] H. Suzuki, H. Ohnishi, and K. Shigemasu. “Goal-directed processes in similarity judgment”. In: *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society (pp. 343-348)* Hillsdale. NJ: Erlbaum (1992) (cit. on p. 6).

[Swe+91] J. A. Swets, D. Getty, R. Pickett, et al. “Enhancing and Evaluating Diagnostic Accuracy”. In: *Medical decision making: an international journal of the Society for Medical Decision Making* 11 (Feb. 1991), pp. 9–18 (cit. on p. 101).

[TR10] K. Taylor and D. Rohrer. “The Effects of Interleaved Practice”. In: *Applied Cognitive Psychology* 24 (Sept. 2010), pp. 837–848 (cit. on p. 16).

[Tor52] W. S. Torgerson. “Multidimensional scaling: I. Theory and method”. In: *Psychometrika* (1952) (cit. on p. 100).

[TW84] M. Treisman and T. Williams. “A theory of criterion setting with an application to sequential dependencies”. In: *Psychological Review* 91 (Jan. 1984), pp. 68–111 (cit. on p. 18).

[WDJ11] C. Wahlheim, J. Dunlosky, and L. Jacoby. “Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging”. In: *Memory & cognition* 39 (July 2011), pp. 750–63 (cit. on pp. 16, 17).

[WFJ12] C. Wahlheim, B. Finn, and L. Jacoby. “Metacognitive judgments of repetition and variability effects in natural concept learning: Evidence for variability neglect”. In: *Memory & cognition* 40 (Jan. 2012), pp. 703–16 (cit. on p. 16).

[WL70] L. Ward and G. Lockhead. “Sequential effects and memory in category judgments”. In: *Journal of Experimental Psychology* 84 (Apr. 1970), pp. 27–34 (cit. on p. 18).

[WJ74] P. Werbos and P. John. “Beyond regression : new tools for prediction and analysis in the behavioral sciences”. In: (Jan. 1974) (cit. on pp. 27, 121).

[Wil13] A. J. Wills. “Models of categorization”. In: *The Oxford handbook of cognitive psychology*. Ed. by In D. Reisberg (Ed.) Oxford University Press, 2013, pp. 346–357 (cit. on p. 25).

[Yan+17] V. Yan, N. Soderstrom, G. Seneviratna, E. Bjotk, and R. Bjork. “How Should Exemplars Be Sequenced in Inductive Learning? Empirical Evidence Versus Learners’ Opinions”. In: *Journal of Experimental Psychology: Applied* 23 (Aug. 2017) (cit. on p. 16).

[Zak+03] S. R. Zaki, R. M. Nosofsky, R. D. Stanton, and A. L. Cohen. “Prototype and Exemplar Accounts of Category Learning and Attentional Allocation: A Reassessment”. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29 (June 2003), pp. 1160–1173 (cit. on p. 37).

[ZH99] S. Zaki and D. Homa. “Concepts and Transformational Knowledge”. In: *Cognitive psychology* 39 (Oct. 1999), pp. 69–115 (cit. on pp. 6, 31).

[ZM09] D. Zeithamova and W. Maddox. “Learning Mode and Exemplar Sequencing in Unsupervised Category Learning”. In: *Journal of experimental psychology. Learning, memory, and cognition* 35 (June 2009), pp. 731–41 (cit. on pp. 15, 29).

[ZSP12] D. Zeithamova, M. Schlichting, and A. Preston. “The hippocampus and inferential reasoning: Building memories to navigate future decisions”. In: *Frontiers in human neuroscience* 6 (Mar. 2012), p. 70 (cit. on p. 29).

[ZM12] D. de Zilva and C. Mitchell. “Effects of exposure on discrimination of similar stimuli and on memory for their unique and common features”. In: *Quarterly journal of experimental psychology (2006)* 65 (Jan. 2012), pp. 1123–38 (cit. on p. 16).

[ZJM11] V. Zotov, M. Jones, and D. Mewhort. “Contrast and assimilation in categorization and exemplar production”. In: *Attention, perception & psychophysics* 73 (Feb. 2011), pp. 621–39 (cit. on pp. 15, 18, 19).

[ZB12] N. Zulkiply and J. Burt. “The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations”. In: *Memory & cognition* 41 (Aug. 2012) (cit. on pp. 16, 17).

[Zul+12] N. Zulkiply, J. Mclean, J. Burt, and D. Bath. “Spacing and induction: Application to exemplars presented as auditory and visual text”. In: *Learning and Instruction* 22 (June 2012), pp. 215–221 (cit. on p. 16).