UNIVERSITÄTSKLINIKUM HAMBURG-EPPENDORF

Institut für Systemische Neurowissenschaften

Direktor: Prof. Dr. med. Christian Büchel

Dissociable Category Learning Systems

Dissertation

zur Erlangung des Doktorgrades Dr. rer. biol. hum.
an der Medizinischen Fakultät der Universität Hamburg

vorgelegt von:

Alina-Nicoleta Dinu
aus Slatina, Olt, Rumänien

Hamburg 2021

Angenommen von der

Medizinischen Fakultät der Universität Hamburg am:

05.11.2021

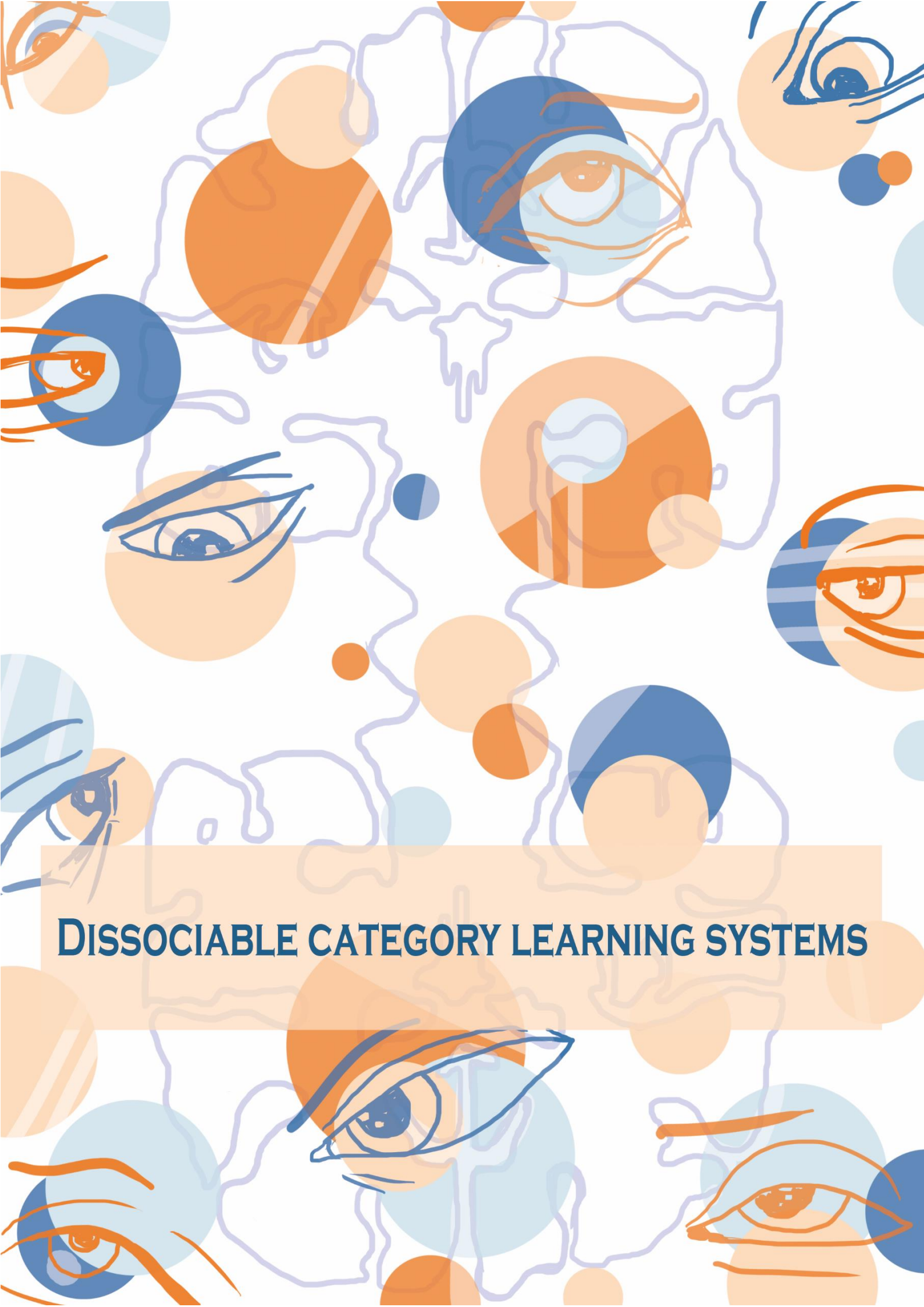
Veröffentlicht mit Genehmigung der

Medizinischen Fakultät der Universität Hamburg.

Prüfungsausschuss, der Vorsitzende: Dr. Jan P. Gläscher

Prüfungsausschuss, zweiter Gutachter: Prof. Dr. Lars Schwabe

Prüfungsausschuss, dritter Gutachter: Priv.-Doz. Dr. med. Gregor Leicht



DISSOCIABLE CATEGORY LEARNING SYSTEMS

To Mom

Pentru mama mea

There is no higher honor that to be told:

'You are your mother's daughter!'

Nu există mândrie mai mare decât să ţi se spună:

„Se vede că eşti fata mamei tale!"

Acknowledgments

First and foremost, I would like to express my gratitude to my supervisors Jan Gläscher and Tobias Sommer. I was one of the few lucky PhD students to have not one, but two enthusiastic, supportive and kind supervisors. Thank you for your suggestions, input and help. Thank you for your patience, for always welcoming my questions and ideas. Without your guidance, I would not have been the researcher I am today.

I would like to extend my gratitude to our collaborator Rene Schlegelmilch for his immense help with computational modeling. I sincerely appreciate your dedication to the project and eagerness to help. I have learned a lot from you.

I would like to acknowledge that my project was part of a collaborative research consortium “Flexible Learning under Stress” funded by the Landesforschungsförderung Hamburg. I would like to thank all members of this consortium for their hard work. I have enjoyed our meetings and engaging discussions and I am looking forward to future developments and collaborations.

I would like to thank Kira Diermann for being the perfect HIWI. You made the recruitment process run so smoothly, helped me with the German instructions and adapted so flexibly to all the technical problems we had in the process. I cannot thank you enough. I would like to also thank our great MR technicians, Katrin Bergholz, Kathrin Wendt and Waldemar Schwarz for their help in collecting the fMRI data. The long scanning days were a little better in your company.

I also had great pleasure of being part of two wonderful research groups: the Gläscher and the Sommer lab. I am thankful for the supportive environment in both groups, for always being able to get valuable feedback on my work and for being able to learn so much from each member and project. Very special thanks go to my dear friend and colleague, Janine Bayer. I am so glad our painting sessions turned in such a beautiful friendship. I will be forever grateful for your help and support during my PhD. I cannot thank you enough for your comments and input on the thesis.

I am glad that I got to share an office with Gina Joue, Saurabh Kumar, Christoph Korn and Tessa Rusch. Our office has always had such a good atmosphere (and most of the time also good snacks!). Thank you for all the uplifting discussions and for being so nice to me. Thank you Gina, for giving the best hugs.

I have also been fortunate to have great colleagues outside my office. I would like to thank Vivien Albrecht for our great conversations and for being so warm and friendly. Our Zoom lunches have made every week a little better. I am grateful that Karita Ojala has recently joined our institute. Thank you for our long dinners and talks. Thank you for all your encouraging messages. Lastly, I would like to thank Alexandros Kastrinogiannis, Alina Koppold, Ben Kupper-Smith, Francisco Lagos Fritz, Lisa Doppelhofer, Mana Ehlers, Riccardo Galli for always being just one knock away from a good laugh, a friendly conversation or a fun pizza evening.

To Emma Jenks, my dear friend, thank you for making Corona times twice as productive. Our Zoom sessions kept me going during these hard times. We both struggled with writing and analyses but having you there always kind, supportive and cheerful has helped immensely.

I would like to thank Cristiana Dimulescu for being my person. Cristiana, you understood better than anyone what a journey this has been for me. I am so thankful that I can always turn to you, that we share the same scientific and personal values. Thank you for all your support and advice throughout these years, for all the proofreading and the stats questions, and all the “auditory system” laughter. Through you, I also got to know Ana Dumitru, who is not only a dear friend but also a wonderful artist who designed the beautiful cover of this thesis. Ana, thank you so so much!

I wish to thank all my friends, spread around all corners of Germany and Europe. I would like to thank Ulrike Nowak, whom I met thanks to the research consortium. Ulrike, thank you for all your wholeheartedly support during this time. Thank you for all the heartwarming post-cards and cups of tea, for all the theater plays and Art Nights. I take great comfort in knowing you “are always by my side”. I would also like to thank my special friend, Alina Dima. Alina, I am so glad our friendship has lasted 16 years, four countries and many confused

documents from German authorities due to our similar names. I cherish that we can always talk as if no time or distance has passed. Special thanks go to my friend Kalika Mehta. Kalika, you are an inspiration to me. You never cease to surprise me with your strength and courage. I am so glad to have met you. Lastly, I would like to thank Cristina Cozari. Cristina, I have told you many years ago that you made Jacobs University so much better. Thank you for that first day we met in the Intro to Cognitive Psychology class and for the hundreds of presentations together that followed. Thank you for all your patience with my slides, for the late night projects and for being the friend and roommate I needed.

Many thanks go to the Maastricht Crew, my dear friends: Julio Rodriguez Larios, Jan Brammer, Koen Frolichs and Till Steinbach. Guys, your enthusiasm for science is contagious. Every time I talk to you I feel recharged. I am so glad that the tough Master's Program led to such a great crew. We all ended up doing PhDs, and although our topics are so different, the joy and passion of discussing them are always there. I am deeply grateful for that.

I would like to thank my family, who even though did not understand what I was doing so far away in cold and gloomy Germany, supported me relentlessly. Mulțumesc tututor celor de acasă, din România, pentru că au crezut tot timpul în mine. Deși sunt departe, am știut mereu că vă gândiți la mine. I especially wish to thank my sister, Cristina Ciocîrlan, for always believing in me. Thank you for being always so caring and understanding, for all the weekend trips and Christmas trees, and for reminding me that I am strong even when I did not believe it. Special thanks go to my youngest supporter and his mom, Andrei and Oana Krista: Mulțumesc din suflet pentru tot ce ați făcut pentru mine! Ați fost mereu modelul meu de urmat". Lastly, I would like to thank the strong "Niță" sisters, Oana and Consuela. Fetelor, thank you for the wonderful trips we have taken these years and for a friendship that lasted two decades. Thank you for listening to me and cheering me up, for always saying "știm noi că poți" even when I did not want to hear it.

Last but not least, I would like to thank my boyfriend, Daniel Hoffmann. Daniel, you have been the best partner one could ask for during a PhD. No words will be enough to express my appreciation. You have been so incredibly patient,

kind and loving. You have made me laugh at the end of some of the darkest days - when the scanner cooling system stopped working for the 10th time, when yet another participant cancelled last minute, when the eye-tracker would just not calibrate and when the eye-tracking data seemed hopeless. You have made the happy days even more cheerful by jumping with me at the news I got the post-doc grant or by baking a done scanning cake. Thank you for listening countless times to all my presentations and for helping me with formatting. Thank you for tirelessly proofreading this work and actually taking the time to understand it even if we are from completely different fields. A complete thank you section to you would probably be as long as this thesis, so I will stop here by saying: Thank you for your love and good food!

Summary

One of the fundamental cognitive operations in which we are engaged every day is the categorization of stimuli into distinct classes. Through it we simplify our immediate environment, structure incoming information and thus reduce cognitive load. Categorization can be done by applying logical rules (e.g. sturdy and waterproof indicates good quality) or by memorizing each individual item (e.g. a specific brand, such as Bosch or Siemens, indicates good quality). These two different categorization mechanisms were investigated in this thesis using a novel choice format adaptation of the type II and type VI problems introduced by Shepard et al. (1961). In type II problems, items can be optimally categorized by applying a two-dimensional eXclusive OR rule (XOR) (e.g. small and circular OR big and triangular belong to Category A). In type VI problems, there is no rule linking items of a category and the optimal solution is memorization. This new adaptation added the much more ecologically valid probabilistic monetary feedback. Moreover, it reduced strategy switch variance through instructions on the optimal strategies. According to these underlying optimal strategies, logical rule finding and memorization, the newly adapted problems were referred to as rule-based and stimulus-based tasks, respectively. The two tasks were compared using behavioral, attentional, cognitive modeling, functional Magnetic Resonance Imaging (fMRI) and model-based fMRI approaches. Rule-based and stimulus-based categorization occurred at the same rate, making this the first study in the literature to find equal learning points in type II and type VI tasks. Striking post-learning behavioral differences were found at the reaction time level, with participants needing twice as much time to categorize in the stimulus-based task than in the rule-based task. Eye-tracking measures revealed differences in attentional allocation suggestive of the underlying optimal strategies. Participants in the rule-based task paid more attention to the relevant rule-forming features than to the irrelevant features. Moreover, attention to features irrelevant to the task decreased steadily as a function of learning, but surprisingly, did not cease once the correct rule had been found. In the stimulus-based task, attentional allocation was minimally altered by learning. The cognitive computations behind the two tasks were investigated using a promising new model, the Category Abstraction Learning

(CAL, Schlegelmilch et al., 2021) whose architecture mirrored the two types of category learning investigated. CAL unraveled that good performance in both tasks requires high encoding and retrieval abilities. Unlike the stimulus based task, the rule-based task is characterized by sharp generalization gradients, reflecting optimal generalization of rule-like information. As expected from previous studies, rule-based categorization relied more heavily on the prefrontal cortex and hippocampus than stimulus-based categorization. By contrast, stimulus-based categorization elicited more striatal and insular involvement than rule-based categorization. Model-based fMRI analyses confirmed that the prefrontal activity was associated with rule prediction. Collectively, this work suggests that rule-based and memorization-based category learning are dissociable systems which optimally interact and inform each other.

Zusammenfassung

Einer der grundlegenden kognitiven Vorgänge mit denen wir uns tagtäglich beschäftigen ist die Kategorisierung von Reizen in verschiedene Klassen. Durch diese vereinfachen wir unsere unmittelbare Umgebung, strukturieren eingehende Informationen und reduzieren so die kognitive Belastung. Kategorisierung kann durch die Anwendung logischer Regeln (z.B. robust und wasserdicht deutet auf gute Qualität hin) oder durch das Einprägen einzelner Gegenstände (z.B. eine bestimmte Marke wie Bosch oder Siemens deutet auf gute Qualität hin) erfolgen. Diese beiden unterschiedlichen Kategorisierungsmechanismen wurden in dieser Arbeit anhand eines neuartigen Auswahlformats untersucht, das eine Adaption der von Shepard et al. (1961) eingeführten Typ-II- und Typ-VI-Probleme darstellt. Bei Typ-II-Problemen können Gegenstände durch Anwendung einer zweidimensionalen XOR-Regel optimal kategorisiert werden (z.B. klein und rund oder groß und dreieckig gehören zur Kategorie A). Bei Typ VI-Problemen gibt es keine Regel die Gegenstände einer Kategorie verbindet, die optimale Lösung ist also das Auswendiglernen. Diese neue Anpassung umfasst das ökologisch validere probabilistische monetäre Feedback. Außerdem reduzierte sie die Varianz in Strategiewechseln durch Anweisungen zu optimalen Strategien. Entsprechend dieser zugrundeliegenden optimalen Strategien, dem Finden logischer Regeln und dem Auswendiglernen, wurden die neu adaptierten Probleme als regelbasierte bzw. stimulusbasierte Aufgaben bezeichnet. Die beiden Aufgaben wurden mit Hilfe verschiedener Ansätze verglichen, die Verhaltens- und Aufmerksamkeitsaufgaben, kognitive Modellierung und funktionelle Magnetresonanztomographie (fMRT) umfassen. Die regelbasierte und die stimulusbasierte Kategorisierung erfolgte mit der gleichen Geschwindigkeit, so dass dies die erste Studie in der Literatur ist, die gleiche Lernpunkte bei Typ-II- und Typ-VI-Aufgaben findet. Auffällige Verhaltensunterschiede nach dem Lernen wurden auf der Ebene der Reaktionszeit gefunden, wobei die Teilnehmer in der reizbasierten Aufgabe doppelt so viel Zeit für die Kategorisierung benötigten wie in der regelbasierten Aufgabe. Eye-Tracking-Messungen zeigten Unterschiede in der Aufmerksamkeitsverteilung, die auf die zugrundeliegenden optimalen Strategien schließen lassen. Die

Teilnehmer:innen der regelbasierten Aufgabe schenkten den relevanten, regelbildenden Merkmalen mehr Aufmerksamkeit als den irrelevanten Merkmalen. Darüber hinaus nahm die Aufmerksamkeit auf die für die Aufgabe irrelevanten Merkmale in Abhängigkeit vom Lernprozess stetig ab. Sie fielen aber überraschenderweise nicht gänzlich weg, sobald die korrekte Regel gefunden worden war. Bei der stimulusbasierten Aufgabe wurde die Aufmerksamkeitsverteilung durch das Lernen nur minimal verändert. Die kognitiven Berechnungen hinter den beiden Aufgaben wurden mit einem vielversprechenden neuen Modell untersucht, dem Category Abstraction Learning (CAL, Schlegelmilch et al., 2021), dessen Architektur die beiden untersuchten Arten des Kategorienlernens widerspiegelt. CAL enthüllte, dass eine gute Leistung in beiden Aufgaben hohe Enkodierungs- und Abruffähigkeiten erfordert. Im Gegensatz zur stimulusbasierten Aufgabe ist die regelbasierte Aufgabe durch scharfe Generalisierungsgradienten gekennzeichnet, die eine optimale Generalisierung von regelbasierten Informationen widerspiegeln. Wie aufgrund früherer Studien zu erwarten, stützte sich die regelbasierte Kategorisierung stärker auf den präfrontalen Kortex und den Hippocampus als die reizbasierte Kategorisierung. Im Gegensatz dazu löste die stimulusbasierte Kategorisierung eine stärkere Beteiligung des Striatums und der Insula aus als die regelbasierte Kategorisierung. Modellbasierte fMRT-Analysen bestätigten, dass die präfrontale Aktivität mit der Regelvorhersage verbunden war. Insgesamt legt diese Arbeit nahe, dass regelbasiertes und auf Auswendiglernen beruhendes Kategorienlernen dissoziierbare Systeme sind, die auf optimale Weise interagieren und sich gegenseitig informieren.

Table of Contents

1. Introduction

Categorization is crucial for survival. Without being able to correctly categorize a food item as edible or poisonous, or an incoming animal as predator or prey, many species would not be able to survive. For humans, categorization is one of the fundamental cognitive operations through which we simplify our immediate environment, structure incoming information and thus reduce cognitive load (Ashby & Valentin, 2017).

Due to its uncontested importance, category learning has been extensively researched in both humans and animals. The current work focuses on one of the multiple branches of human categorization research, namely perceptual categorization of visual stimuli (Ashby & Maddox, 2005). Formally, this type of categorization (also referred to as classification) is defined as the process of placing items into classes or groups based on shared characteristics which results in "the act of responding the same to all members of one stimulus class and differently to members of other classes" (Cantwell et al., 2017, p. 32).

Depending on the nature of the classification problem of interest, perceptual categorization paradigms mainly employ three different types of stimuli: stimuli with separable dimensions and discrete features (i.e. shape: square, triangle, circle), stimuli with separable dimensions and continuous features (varying length on a continuous scale, Gabor patches) and random dot patterns.

This project used the first class of stimuli with separable dimensions and discrete features, and was based on the most influential study to employ these type of stimuli: Shepard et al. (1961). By using stimuli with three dimensions and binary features, the authors introduced six iconic types of category learning problems. **Figure 1** depicts six rectangles, each containing one problem type. Within each rectangle, the left and right side correspond to category A and B, respectively. In type I problems, the two categories can be easily identified by applying a uni-dimensional rule (black figures belong to category A). The type II problem entitles a two-dimensional eXclusive OR (XOR) rule application (stimuli that are black and small OR white and large belong to category A, shape is irrelevant). The next three problem types (III, IV and V) follow a rule-plus-

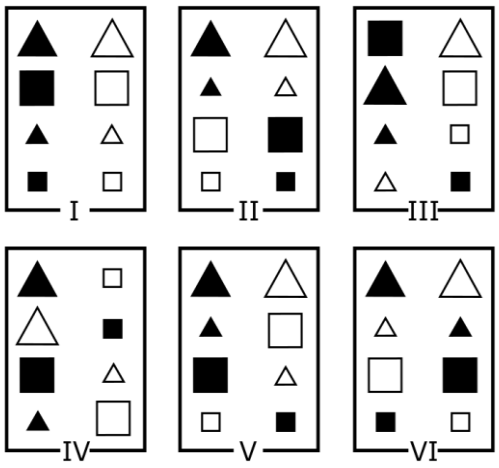


Figure 1*. Shepard et al. (1961)'s category learning problems. Each rectangle corresponds to one problem type. Within each rectangle, the stimuli on the left side belong to Category A and the ones on the right side belong to Category B. Figure adapted without permission from the original paper.*

exception structure - most members of a category follow a uni-dimensional rule (i.e. type III black stimuli are category A) with one exception (i.e. the white triangle is also category A – type III). The last problem type, type VI, has no rule-like solution, thus requires stimulus-category label memorization.

Shepard et al. (1961) showed that when it comes to ease of acquisition, the six problem types can be ordered as follows: Type I < Type II < Types III, IV, V < Type VI from easiest to most difficult. The impact of this ordering and the problems themselves on categorization literature has been tremendous. It has led to a redefinition of concept learning, innovation of unsupervised learning (Love, 2002) and re-evaluation of the assumption of independence in stimuli properties (Love & Markman, 2003). Particularly in the field of cognitive modeling, the six problems have always been the skeleton on which all research was built; every new category learning model (e.g. ALCOVE (Kruschke, 1993), SUSTAIN (Love et al., 2004), DIVA (Kurtz, 2007)) had to first and foremost be able to capture Shepard et al. (1961)'s ordering. Noteworthy, despite the numerous stimulus sets used to assess, investigate, and apply these problems (e.g. geometric forms, algae, instruments, beetles, flowers), the original task format has been kept throughout the decades. That is, participants are presented with one stimulus at a time and learn with the help of feedback

(immediate or delayed) which of the two possible categories the stimulus belongs to.

The current work aimed to advance the standard paradigm, in which participants have to identify a stimulus' category label, to a format in which participants have to *choose* out of two simultaneously presented stimuli, the stimulus belonging to a target category (e.g. category A). Introducing the element of choice in these problems is not only a meaningful methodological advancement (Wang & Ashby, 2020), but also a natural transition towards a higher ecological validity. In real life, categorization and choice often co-occur, such as when choosing what type of food to order, what type of fruits to eat or what genre of book to read. Hence, it is essential that new experimental paradigms try to incorporate this aspect.

This project implemented two-alternative choice categorization in Shepard et al. (1961)'s type II and type VI problems. To reiterate, a type II problem can be correctly solved by the application of a disjunctive *rule* (e.g. fruits that are small and yellow or big and red are poisonous) whereas in type VI the solution requires *memorization of* each stimulus and its category label (e.g. ivy berries and holly berries are poisonous).

The two problems were selected in an attempt to disentangle two category learning systems: declarative and procedural. The idea of distinct categorization systems was first postulated by Ashby et al. (1998) in the framework of the Competition between Verbal and Implicit Memory Systems (COVIS) model. The model posits that there are two distinct competing category learning systems: a declarative system, encompassing explicit reasoning through clearly verbalizable rules and a procedural system, characterized by implicit reasoning which cannot be verbalized. The model distinguished itself from contemporary models by assuming that these two systems are also clearly separable at a neural level. That is, the declarative system recruits the prefrontal and cingulate cortices while the procedural system is striatal-based.

Over the last two decades, the existence of two separate instead of a unitary system of categorization has been intensely debated (Ashby & Maddox, 2011; Cantwell et al., 2017; Edmunds et al., 2018; Filoteo et al., 2001; Filoteo et al., 2005; Maddox & Ashby, 2004; Nomura et al., 2007; Nosofsky & Kruschke,

2002). Most studies challenging the COVIS model have done so by comparing rule-based and information-integration paradigms (Ashby & Maddox, 2011; Ashby, Smith & Rosedahl, 2019). Typically, these paradigms use large stimulus sets with discrete dimensions and continuous features (e.g. sinusoidal gratings with varying spatial frequency and orientation). While these paradigms do differ in terms of stimuli, category and task structure from Shepard et al. (1961)'s problems, the core assumptions are conceptually similar. In this framework, rule-based tasks are described as tasks in which the optimal solution is a clearly verbalizable logical rule – akin to type II problems that require an XOR rule. Information-integration tasks on the other hand, are tasks in which the optimal strategy involves the integration of different stimulus information at a pre-decisional stage. Since this integration process cannot be easily verbalized, it is regarded as implicit. It can be argued that if performed using a memorization strategy, Shepard et al. (1961)'s type VI problem also entails a non-verbalizable solution, and thus the problem resembles an information-integration task. Therefore, the type VI problem can be used as an equivalent for procedural learning.

Although, at a behavioral level, the literature is leaning towards a consensus that these systems are indeed dissociable (but see Stanton & Nosofsky, 2007, 2013), the discussion regarding the neural level is far from being settled. This thesis contributes to solving this debate by comparing neural activity in the two-choice adaptation of type II and type VI problems using functional Magnetic Resonance Imaging (fMRI). The neural signatures were examined in greater detail with the help of model-based fMRI using the newly developed *category abstraction learning model* (CAL) (Schlegelmilch et al., 2021, under review). This model is particularly promising in addressing the type II / type VI differences, since its internal architecture contains two separate networks mirroring the type of learning investigated: a *rule network* and a *configural memory network*. The *rule network* solves the task via rule prediction and is based on how strongly an observed stimulus feature is associated with the available categories. The *configural memory network* solves the task by "memorization" and is based on a recall heuristic, which only activates the instance from memory that is most similar to the presented stimulus.

Lastly, the present work also addressed the type II / type VI comparison from an attentional angle. It was aimed that, by using a well-established proxy for overt attention, namely fixation counts (Liversedge & Findlay, 2000), the differences between the two problem types would be understood in greater detail. This is an important addition to the categorization literature since until now there has been only one study to systematically assess attentional allocation in Shepard et al. (1961)'s problems, the eye-tracking study by Rehder and Hoffman (2005). The authors demonstrated for the first time what had only been presumed for four decades, namely that participants allocate attention optimally based on the problem type. Moreover, it was found that participants started each problem type by fixating evenly on all three dimensions. It was only *after* learning that they abruptly shifted their fixations to the relevant dimensions. This observation challenged crucial assumptions of contemporary models such as the gradual learning predicted by ALCOVE (Kruschke, 1992) or the hypothesis testing mechanisms incorporated in RULEX (Erickson & Kruschke, 1998). Thus, it is evident that eye-tracking measures can have a powerful impact not only on understanding the problems per se but also on current and future models of categorization. Consequently, this study revived eye-tracking research on Shepard et al. (1961)'s problems which was deemed long overdue considering that most categorization models have attention as a key component of their system.

In order to clearly highlight their underlying strategies and to facilitate recognition, the compared problem types will henceforth be referred to as rule-based (type II) and stimulus-based (type VI). When discussing previous work on the problems they will still be referred to by their original terminology as type II and type VI. The problems themselves as a collective will be referred to as either Shepard et al. (1961)'s problems or Shepard's problems interchangeably.

To sum up, this work sets out to adapt these two problems to a novel choice format and explore the associated behavioral and attentional mechanisms. Moreover, the goal was to gain a deeper understanding of the neural mechanisms of rule-based and stimulus-based category learning and to shed light on the ongoing debate on whether categorization involves two distinct neural systems.

This thesis continues with a presentation of past applications of Shepard's problems followed by a detailed description of the rationale and structure of the newly developed paradigm (**Chapter 2**). The next three chapters contain empirical work assessing this paradigm. **Chapter 3** covers the behavioral correlates of the rule-based and stimulus-based task. **Chapter 4** presents eye-tracking assessments of attentional mechanisms elicited during the two tasks. In **Chapter 5** the cognitive model CAL is introduced and applied to the behavioral data. **Chapter 6** displays the neural correlates of rule-based and stimulus-based categorization using model-free and model-based fMRI. Each empirical work chapter contains individual introduction, material and methods, results and discussion sections. In **Chapter 3**, **Chapter 4,** and **Chapter 6,** the research hypotheses are presented in the introductory sections. **Chapter 5** contains two additional sections, one describing mechanisms of the CAL model and one presenting challenges for the current paradigm. **Chapter 7** concludes the thesis and presents suggestions for future research.

2. A Novel Two-Way Categorization Paradigm

As mentioned in the introductory chapter, this project attempted to advance the standard Shepard et al. (1961) type II and type VI problems to the more ecologically valid choice format. This chapter briefly summarizes past adaptations of these problems restricting the literature to human studies that included both type II and type VI problems. It has to be acknowledged that important work has also been done using type II in combination with other types such as type I (Mack et al., 2016) and type IV (Kurtz et al., 2013; Love & Markman, 2003). The summary is followed by a section highlighting the novelties of the current paradigm and a detailed description of the paradigm.

2.1 Shepard's cube

Before discussing the original implementation, it is noteworthy that in addition to introducing the six iconic categorization problems, Shepard et al. (1961) also presented a way of visualizing the stimuli to facilitate comparison among different stimulus sets. The authors suggested that the three dimensions of the stimuli can be regarded as three dimensions of a cube, and each dimension's binary feature can be thought of as vertex. **Figure 2** depicts this cubic illustration for the type I, type II and type VI problems. The four stimuli with grey contour are assigned to one category (equivalent to the left side of each

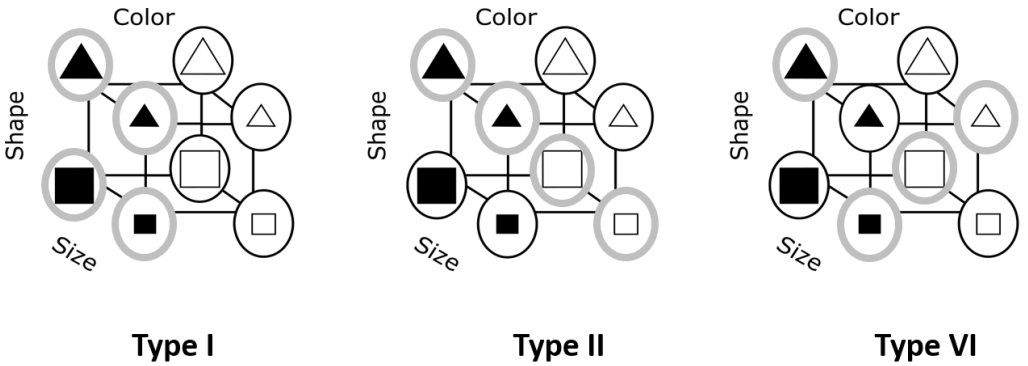


Figure 2. *Cubic representation of Type I, Type II and Type VI problems. The eight stimuli are displayed on a cube by assigning each of their varying dimensions to one of the cube's dimensions. On any of the 12 faces of the cube the four stimuli share one feature. In each cube, the four highlighted stimuli belong to one category (i.e. category A) and the remaining ones belong to the remaining category (i.e. category B). Figure adapted without permission from Shepard et al. (1961).*

rectangle in **Figure 1**) and the four without grey contour to the other category (equivalent to the right side of each rectangle in **Figure 1**). Interestingly, this depiction is also representative of the problems' (assumed) complexity level. In type I, all members of a category share a face of a cube or plane which translates into a low level of complexity. In type II, members of a category are no longer located on a face, but on a hyperplane which "cuts" the cube diagonally, indicating an increase in complexity with respect to the previous problem. In type VI, a category cannot be captured by either a plane or a hyperplane, and thus it has the highest level of complexity. Importantly, all Shepard et al. (1961) replication studies and studies using any subset of these problems (the current work included) follow this cubic visualization when constructing and presenting their stimuli.

2.2 Previous paradigms

This section contains an extensive description of the previous work using Shepard's problems. The following studies are covered: Shepard et al. (1961), Nosofsky et al., (1994), Love (2002), Smith et al. (2004), Rehder and Hoffman (2005) and Lewandowsky (2011). For each study, aspects such as the sample size, problems used, experimental paradigm and instruction type are covered in great detail. This summary is intended as a support for the future researchers interested in conducting new studies on these problems, replication studies or reviews. The details from these studies that are crucial for the current thesis are summarized at the beginning of the next section.

Shepard et al. (1961) used squares and triangles for concept illustration only and more complex stimuli for the actual tasks. These stimuli had a triangular format with each vertex representing one dimension. The binary features of each dimensions consisted of two thematically related objects. The paper provides examples of these objects (e.g. candle and light bulb, violin and trumpet), but the full stimulus set is not described. The two category labels themselves were not thematically meaningful (e.g. furniture versus instruments), and instead were assigned to be either letters, symbols or numbers. All participants performed type I, II and VI problems and either of the type III, type IV, or type VI problem in a randomized order. Each problem was performed five

consecutive times, and for each repetition a different stimulus set was used (resulting in a total of five stimulus sets for the whole experiment). Before the start of the task, participants were made aware of the stimulus construction. With respect to how to perform, they were only told that there will be different types of problems with similar difficulty and the beginning of a new problem type will be announced beforehand. The stimuli were presented sequentially, and each choice was followed by immediate auditory feedback on its correctness. Each repetition stopped when participants achieved 32 consecutive classifications. Participants completed the problems in one hour sessions, three times a week. The total number of sessions differed depending on individual performance. After successful completion of each problem type, subjects were asked to describe what strategies they used. No suggestions about the existence of any rule were made. All in all, six participants were tested.

Nosofsky et al., (1994) were the first to undergo a replication study with a larger sample of 120 participants. The authors made three changes to the initial study design. First, the three stimuli dimensions used (shape, size, and types of lines inside the shape) were no longer spatially separated. Second, their associated categories were simply labeled as "1" or "2" .Third, each participant performed only two out of the six problem types, counterbalanced such that all problems were equally represented. Each problem stopped when participants completed 32 consecutive correct classifications (as in the original paper) but an additional constraint was set that all problems would stop after a maximum of 400 trials. Subjects received rule-like instruction by being told that the "relevant rule and dimensions for the second problem were chosen independently of those that were relevant in the first problem" (Nosofsky et al., 1994, p. 355). The task format was a standard sequential stimulus presentation with immediate feedback provided upon choice. Neither feedback type nor feedback duration were mentioned.

The replication by Love (2002) restricted the problem set to type I, type II, type IV and type VI with each participant completing only one problem type. The sample size was doubled from the previous study (252 participants). This time the stimuli' varying dimensions were randomized: size (small or large), color (blue or purple), texture (smooth or dotted) or a diagonal cross (present

or absent). Moreover, instead of simply labeling categories (e.g. category 1 or category 2), a fourth stimulus feature, border color, acted as a label. Participants were instructed that in each trial they have to predict the border color of the presented stimulus, and that this color depends on the value of the other dimensions. The feedback consisted of a tone (positive if correct, negative if incorrect) together with the stimulus surrounded by the right border. The complete stimulus was kept on the screen for 1.5 seconds. A 1 second break concluded each trial. Irrespective of their performance, the task stopped after 80 trials.

Smith et al. (2004) returned to each of the 47 participants performing all problem types in a randomized sequential order. The stimuli had the standard three possible varying dimensions and resembled those used by Shepard et al. (1961) to introduce the concepts and the cube (with the only difference that the colors could be either dark gray and white rather than black and white). The two categories were labeled as category 1 and category 2. Participants were instructed that once they learn "the right category or rules they will work for the whole task" (Smith et al., 2004, p. 403), and were announced before the task (problem type) had changed. Interestingly, this was the first study in which participants were motivated to perform correctly by receiving points for each correct classification (no points were deducted for incorrect classification). Additionally, they were told that the performers with the highest number of points can receive a prize of up to 20 dollars.

The study by Rehder and Hoffman (2005) returned to assessing only type I, type II, type IV and type VI problems. Each participant was assigned to one type only, resulting in 18 participants per problem (72 participants in total). Due to the eye-tracking nature of the study, the stimuli had spatially separable dimensions. The three dimensions were varying text symbols: $ and ¢, ? and !, + and - .The two possible categories were labeled as "red" or "green". Each stimulus classification was followed by auditory feedback (different tones for correct and incorrect) and the stimulus remained on the screen for a period of 4 seconds after feedback. The task stopped after 32 consecutive correct trials. After each block of 8 trials participants were informed how close they were to this criterion. Each participant was allocated a maximum of 224 trials to reach

this goal. The authors do not mention how the participants were instructed, but from the discussion, it can be deduced that no rule-like instructions were given.

The latest study that systematically assessed all problem types was Lewandowsky (2011). All participants performed all problem types in a pseudo counterbalanced order (i.e. half of the participants started with type I, II, III, the other half with type VI, V, and VI). The eight stimuli differed in color (unfilled or red), shape (square or circle), and size (small or large) and were assigned to either category 1 or 0. Participants performed a maximum of 192 trials, with each problem being terminated upon 32 consecutive correct trials. Each trial was followed by visual feedback containing only the word “correct” or “wrong” for a period of 2 seconds. With respect to instruction type, the authors only mention that participants were aware of the change in problem type before the start of a new type.

Lastly, a recent fMRI study by Mack et al. (2020) used the type I, type II and type VI problems. The stimuli were more visually complex than those in the past studies, and consisted of beetles with the three varying dimensions being antenna size (thick or thin), leg size (thick or thin) and mouth type (shower or pincer). On each problem type the two stimuli were labeled as either coming from an eastern or western hemisphere, liking cold or warmed temperatures or living in a rural or urban environment. The subjects were instructed that the category assignment was arbitrary. Before the main task, participants were familiarized with the stimuli and the task structure through training tasks. Moreover, the varying dimensions of the stimuli were highlighted and it was emphasized that membership to a category is given by one or more stimulus features. To facilitate fMRI data acquisition, each beetle was presented on the screen for 3.5 seconds. For acquisition purposes, the feedback was delayed by a certain time period which varied randomly between 0.5 and 4.5 seconds. The feedback consisted of an image containing the stimulus and the correctness of the response given, and was displayed for a duration of 2 seconds.

2.3 Methodological highlights

Important methodological observations can be drawn from the summary provided in the previous sections. First, throughout the literature, there have been inconsistencies in the instructions participants received prior to performing the problems, especially with respect to whether or not rule-like solutions are to be expected. These inconsistencies could have fundamentally altered the way subjects learned during the task and can hinder comparison across studies. Second, in most studies subjects were not questioned on how they performed the task, and therefore their strategies are unclear. Third, in some of the studies the same stimulus set was used for multiple problem types, thus resulting in subjects having to undergo a process of stimulus remapping in addition to categorization. Fourth, the feedback modality differed across studies. In the studies after 2000 (with the exception of Lewandowsky, 2011), the feedback displayed the stimulus to-be categorized together with the correct label for a fixed period of time. This aspect alone could have led to a faster learning. Lastly, the studies alternate between within-subject design and between-subject design which could influence the extent of the effects observed. The current work developed a paradigm that, in addition to the two-choice format, resolves these methodological inconsistencies.

A within-subject design was used such that all participants performed both tasks, each task on a separate day. It was ensured that the participants received clear instructions on how to optimally perform *both* types of problems. In addition to eliminating possible sources of variation due to strategy search, this is an important methodological advancement since no study up to date provided clear strategy instructions on the type VI problem. Although previous research has looked into the effect of instruction on type II performance (Kurtz et al., 2013), the effect of instructions on type VI is unstudied. To make sure that the instructions were followed in the rule-based task, participants were asked multiple times throughout the task to report the rule they were using. In the stimulus-based task, the subjects were extensively trained and encouraged to use a memorization strategy.

To counteract possible stimulus-remapping concerns, each problem was performed with a different stimulus set, and the assignment of stimulus sets

to problems was balanced across subjects. Each problem type was performed only once. With respect to feedback, the feedback screen was displayed for a fixed period of time, and consisted of a presentation of the stimulus together with its label. Furthermore, as in Smith et al. (2004), to encourage successful learning and concentration throughout the entire task, participants received points for each successful classification, and were told that these points will lead to more financial compensation at the end of the task.

With these potential confounds addressed, the current paradigm also set to explore two departures from the standard tasks, namely probabilistic feedback and category relabeling, whose rationale is highlighted below.

As far as probabilistic feedback is concerned, up to now no study in the category learning literature employed probabilistic feedback in Shepard's problems. From a methodological point of view, probabilistic feedback is advantageous because it slows down learning, and thereby ensures that there are enough incorrect trials for behavioral analyses and cognitive modeling. From a conceptual point of view, the probabilistic feedback brings more ecological validity to the paradigm. On a daily basis, individuals do not always receive feedback contingent to their classification (e.g. doing good work on a project and receiving negative feedback from a colleague – resulting into questioning whether the right category for the work done is "good"), and have to go through a couple of trials to find the correct category (e.g. presenting the work to multiple unbiased colleagues who give positive feedback, thus confirming that the work can indeed be classified as "good"). All in all, these aspects were considered reasonable motivation to introduce probabilistic feedback into the new paradigm.

With respect to category labels, the current paradigm stepped aside from the conventional category 1 versus category 2 labeling by referring to the two classes as valuable or not valuable. This relabeling had both conceptual and methodological motives. Conceptually, it was yet another small step towards ecological validity since on a daily basis one would rarely be asked to classify incoming objects as 1 or 2. Although this is by far not the first attempt to make category labels less abstract (e.g. Mack et al., 2020), it is the first to take on a value-based angle. This angle has more real-life applicability because most of

the objects in daily life are intermixed with their subjective value (expensive phones, cars, cheap clothes, food, furniture etc.).

From a methodological point of view, the paradigm has potential important implications for the decision making community. One could argue, and perhaps rightfully so, that by displaying two stimuli on the screen and assigning value to each of them, the current paradigm is essentially no different from a standard, extensively-researched reinforcement learning task (Pessiglione et al., 2006). However, in standard reinforcement learning tasks stimulus' features are clearly either predictive or not predictive of reward (e.g. small stimuli have a higher reward probability). While this criterion holds for the current adaptation of the type VI problem (stimulus-based task), it fails for the adaptation of the type II task (rule-based task). Here, the same features (e.g. small) can be both predictive and not predictive of reward (e.g. stimuli with a small size and a triangular shape are valuable but stimuli with a small size and quadratic shape are not valuable). Thus, the current rule-based task opens up new doors for the decision making community, in particular for reinforcement learning models.

To summarize, the experimental paradigm applied in this study has five important aspects: two-stimulus display presentation, strategy-based instructions, within-subject design, probabilistic feedback with the feedback display presenting the stimulus and its associated category (in fact this occurs indirectly for both stimuli presented) and category relabeling with a value angle. These features have the potential to open up new exploratory paths for the category learning research and even lead to advancement in the reinforcement learning literature.

2.4 Paradigm description

2.4.1 Stimuli

Two sets of eight stimuli were created for the two main tasks. The stimuli were visually complex and depicted robot-like figures (humanoids) (**Figure 3A**) and butterflies (**Figure 3B**). In both sets, the stimuli differed in three dimensions: color, number and orientation, each with binary features. The humanoids had

green or blue hats, one or two stars on the belly and shoes pointing upwards or downwards. The butterflies had antennas pointing upwards or downwards, one or two circles on their wings, and red or blue tails. In both stimulus sets, the varying dimensions were equidistant from each other and measured approximately 3 cm².

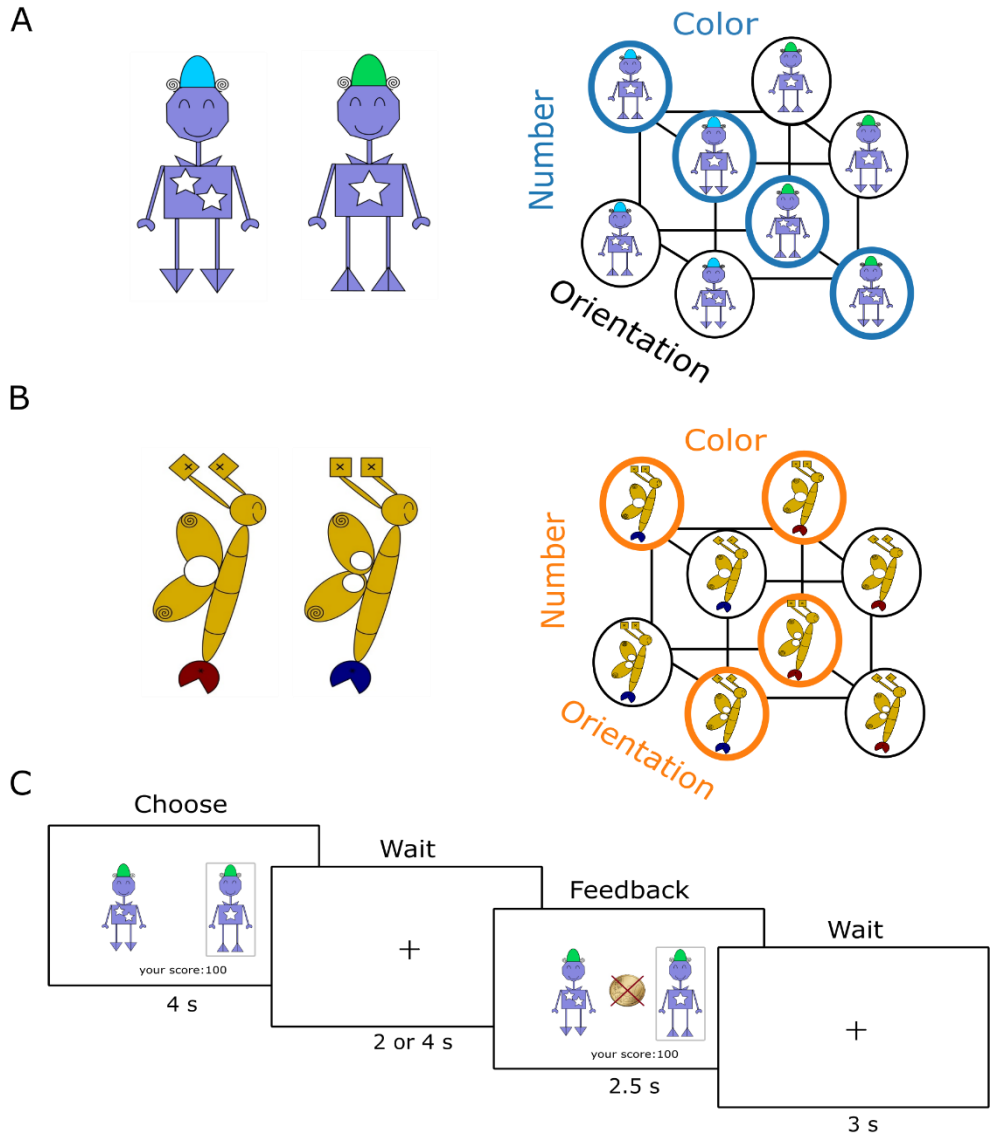


Figure 3. *Tasks description.* ***A.*** *Representative stimulus pair (left) and category membership structure (right) for the rule-based task. The two rule-forming dimensions, Color and Number, are highlighted in blue.* ***B.*** *Representative stimulus pair (left) and category membership (right) for the stimulus-based task. Color, Number and Orientation are written in orange to emphasize that all dimensions are relevant. In both A and B, the four stimuli in the cube circled in the corresponding task color are the four valuable stimuli.* ***C.*** *Example trial. Below each screen its presentation duration is displayed in seconds.*

2.4.2 Task structure

Figure 3C displays an example trial. A trial started with a presentation of a two-stimulus pair. At the bottom of the screen the participants could see their cumulative score. Participants had 4 seconds to choose the valuable stimulus using the left or right arrow key. The chosen stimulus was surrounded by a grey rectangle. The two-stimulus pair did *not* disappear upon choice. The *choice* screen was followed by a *wait* screen displaying a black fixation cross on a white background. After a period of 2 or 4 seconds (randomized across the task), the *feedback* screen appeared. This screen was displayed for 2.5 seconds and contained: the two stimuli, with the selected one highlighted, a 50 cents coin indicating if the choice was correct (normal coin: correct, coin crossed by a red X: incorrect) and the updated cumulative score. If the participants failed to answer within the allotted time, the feedback screen was blank and only displayed the sentence “Please answer faster!”. The trial ended with a *wait* screen presented for 3 seconds. Irrespective of their performance, participants had to complete 160 trials. Throughout the whole task, each choice was followed by probabilistic feedback which was pseudorandomized according to the two-stimulus pair structure.

2.4.3 Pair structure

Four stimuli were assigned to the valuable category and four to the not valuable category. 16 unique two-stimulus pairs were constructed such that in each pair one stimulus was valuable and the other one was not valuable. Care was taken that the stimuli within a pair differed in more than one feature so that the participant cannot completely rule out a feature based on feedback. In other words, there was no pair in which all but one feature was kept constant (same color of the hat, same shoes orientation but different number of starts on the belly). Each individual stimulus was used in four pairs. A pair formed of stimulus 1 on the left side and stimulus 8 on the right side was considered different from a pair with stimulus 8 on the left side and stimulus 1 on the right side. The 16 pairs were presented 10 times with a minimum of one trial gap before the same pair was presented again. The probabilistic feedback was pseudorandomized such that in 2 out of the 10 repetitions (20 % of the cases) each pair was

associated with the *misleading* feedback (explained in detail in **Chapter 3.3.2.5**).

2.4.4 Rule-Based Task

One set of stimuli was randomly chosen for this task (e.g.: **Figure 3A** left). Similar to Shepard et al. (1961), the stimuli were represented in a cubic perceptual space, each dimension corresponding to one dimension of the cube. As before, four stimuli were assigned to the valuable category (i.e. the ones with a blue contour in **Figure 3A** left) and the remaining four were assigned to the not valuable category (i.e. the ones with a black contour in **Figure 3A** left). As in the original Type II problem, the stimuli could be correctly classified by using a disjunctive rule XOR rule (i.e. humanoids with blue hats and one star on the belly are valuable or humanoids with green hats and two starts on the belly are valuable; **Figure 3A** right). All in all, based on the rule-forming dimensions, there were three possible conditions: color and orientation relevant, number irrelevant; color and number relevant, orientation irrelevant; number and orientation relevant, color irrelevant.

Prior to the task, the nature of the stimuli (humanoids or butterflies) and their varying dimensions (color, number, orientation) were explained. Moreover, participants were clearly instructed that the task entails rule application. To further encourage rule-learning, they were told that they will have to report the rule multiple times throughout the task.

2.4.5 Stimulus-Based Task

The remaining set of stimuli was used for this task (i.e. if the rule-based task used the humanoids stimulus set, the stimulus-based task used the butterflies stimulus set; **Figure 3B** left). The four valuable stimuli were selected according to Shepard et al. (1961)'s Type VI problem to not have any rule-like connection between them. This resulted in two possible conditions depending on which four stimuli were assigned to the valuable category (i.e. one in which the stimuli with the orange contour were valuable and one in which the stimuli with the black contour were valuable; **Figure 3B** right).

As in the rule-based task, participants were clearly instructed on the nature of the stimuli (humanoids or butterflies). Additionally, they were instructed on the optimal strategy to perform the task, namely stimulus-category memorization, and were encouraged to use it.

3. Dissociable Behavioral Signatures

3.1 Introduction and Hypotheses

Over the past decades, almost all studies have successfully replicated Shepard et al. (1961) 's initial ordering. To reiterate, the authors found that in terms of difficulty, the six problems can be ordered as follows: type I < type II < type III, type IV, type V < type VI. **Figure 4** displays a summary by Smith et al. (2004) of the findings of four of these studies. It can be seen that, although the overall number of errors has fluctuated across studies, the pattern has been relatively stable. More importantly for the current work, type VI performance has consistently fallen behind type II performance. It has to be mentioned that even in the study Lewandowsky (2011) in which the full ordering was not replicated (due to performance in type II task resembling the ones in type III, IV and V), type II was still found to be by far less difficult than type VI.

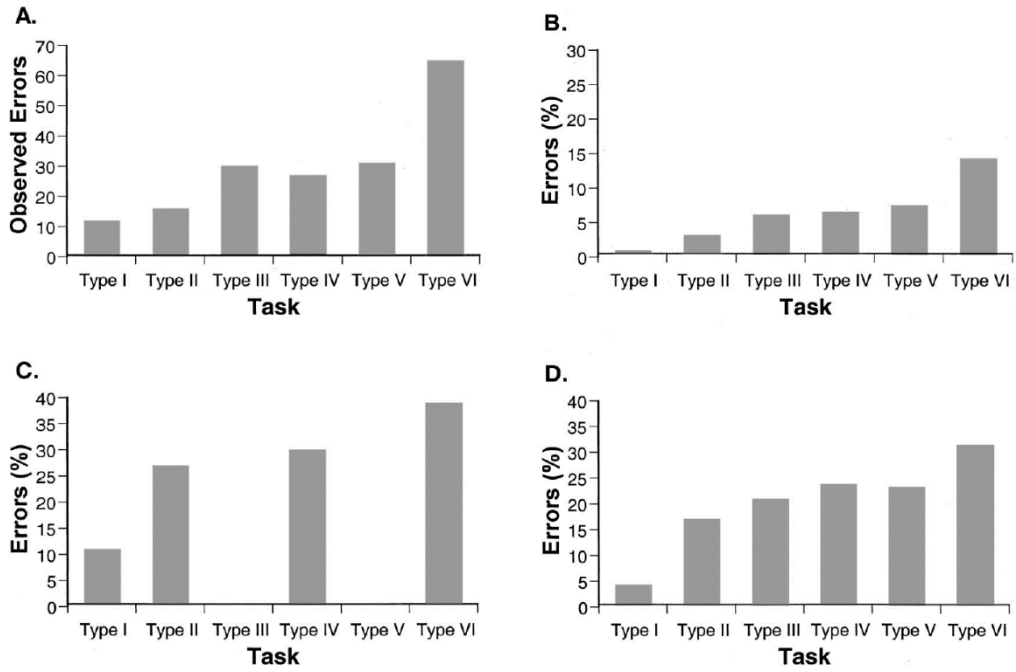


Figure 4*. Performance summary of four replication studies of Shepard et al. (1961).* ***A****. Performance in the original study.* ***B.*** *Performance in Nosofsky et al. (1994).* ***C.*** *performance in Love (2002).* ***D.*** *Performance in Smith et al. (2004). Figure retrieved, caption adapted without permission from Smith et al. (2004).*

The category learning literature in general and the modeling literature in particular, have been focusing so far on capturing the "type II performance advantage", mostly with respect to the type IV problems (Kurtz et al., 2013). The current work departed from this avenue and underwent a detailed comparison of the type II and type VI problems. Specifically, it was aimed to explore for the first time how the two problems behave in a two-alternative choice format. The interest in comparing single-stimulus display to paired display has just begun to arise in the categorization literature. A recent study by Wang and Ashby (2020) assessed differences between single and two-stimulus pair displays in unstructured categorization (i.e. members of a category cannot be found by either rule-learning or memorization), and found a performance advantage when using the latter. In line of this evidence, similar effects would have been expected in the current study. However, the present paradigm also included probabilistic feedback, which is well-known to slow down learning. Thus, any two-stimulus display advantage could have been overridden by the "misleading" feedback.

Up until now, all studies addressing the type II and type VI problems have been concentrating on accuracy measures. To date, only the study by Love (2002) reported reaction times, but these were not analyzed further. Interestingly, despite their large difference in accuracy, the two problems had similar mean reaction times ($M_{type\ II}$ = 1.68 seconds, $SE_{type\ II}$ = 0.97, $M_{type\ IV}$ = 1.69, $SE_{type\ IV}$ = 0.93). This aspect could indicate previously unaddressed similarities between the two problems, and therefore called for further investigation. Consequently, in addition to accuracy measures, this work took a closer look at reaction times and their evolution across the two tasks.

The accuracy and reaction time data were used to assess the following hypotheses:

H1: Regardless of problem type, there will be a difference in the overall numbers of errors between the current adaptation and past adaptations of Shepard's problems.

H2: More errors will be made in the stimulus-based task than in the rule-based task.

H3: The rule-based task will be learned faster than the stimulus-based task.

H4: The rule-based task and the stimulus-based task will not differ with respect to reaction time.

3.2 Materials and Methods

3.3.1 Participants

The study was approved by the local Ethics Committee of the Hamburg Medical Association (ethics number PV5947). Participants were screened for history of psychiatric or neurological disorders and current use of psychoactive medications. For eye-tracking reasons, participants with diopters below -4 or above +4 were not invited to take part in the study. Out of the 53 tested participants, only those who successfully completed the task were included in the analyses. Successful performance was defined in the rule-based task as being able to report the correct rule at the end of the task and in the stimulus-based task as minimum of 20 consecutive correct trials. This resulted in a sample of 30 participants (20 females, 10 males, M_{age} = 26.26, SD_{age} = 2.81).

3.3.2 Training tasks

3.3.2.1 Stimuli

Two sets of eight stimuli were created. Each set consisted of geometrical figures with three dimensions. The first set contained figures which varied in color (red or blue), shape (square or triangle) and filling (filled or not filled) (**Figure 5A**). The second stimulus set contained figures with varying color (green or purple), shape (star or pentagon) and contour (with or without contour) (**Figure 5B**).

3.3.2.2 Task Structure

The structure of the training tasks only differed from the one of the main tasks in timing and numbers of trials performed (**Figure 5C**). In each trial participants were asked to identify the valuable figure within 3.5 seconds, waited

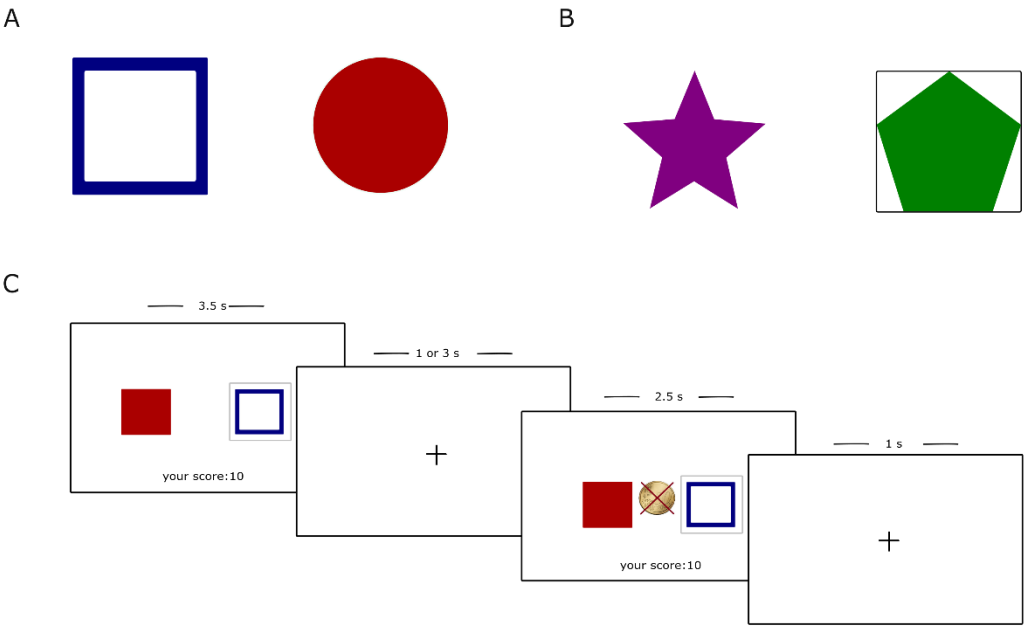


***Figure 5**. Deterministic training tasks. **A.** Representative pair for the deterministic rule-based training task. Stimuli differ in color (red or blue), shape (circle or square) and filling (filled or not filled). **B.** Representative pair for the deterministic stimulus-based task. Stimuli differ in color (green or purple), shape (pentagon or star) and contour (with or without). **C.** Example trial. The numbers on top of each screen indicate its presentation duration in seconds.*

for the feedback for a random period between 1 and 3 seconds, and had only a 1 second inter-trial break. The task ended when the participants completed 15 correct consecutive trials, i.e. correctly selecting a valuable stimulus 15 times.

The underlying category structure of the training task depended on the type of learning investigated in the main task, namely rule-based or stimulus-based. Thus, two training task versions were created: deterministic rule-based training task and deterministic stimulus-based training task.

3.3.2.3 Deterministic Rule-Based Training Task

One of the two sets of stimuli was randomly chosen. As in the main task, category assignments were done based on a disjunctive rule (XOR) (**Figure 1**, type II). In other words, the four valuable stimuli could be identified by using a combination of two dimensions (i.e. color and shape). For example, for the stimulus set in **Figure 5A,** the disjunctive rule was: red circles and blue squares or blue circles and red squares.

Before performing the task, participants were instructed that they can correctly identify the valuable stimuli by applying a rule. In addition, they were aware that depending on the stimulus set assigned, the filling or the contour dimension was irrelevant. At the end of the task, the participants had to report if they found the underlying rule and were asked to describe it.

3.3.2.4 Deterministic Stimulus-Based Training Task

The remaining stimulus set was used (i.e. if the deterministic rule-based task used the set in **Figure 5A**, then the stimulus-based versions used the set in **Figure 5B**). As in the main task, the category structure was constructed such that it could not be predicted by any rule or pattern (**Figure 1**, type VI). At the end of the task, participants were presented with all eight stimuli and had to indicate for each one of them whether it was a valuable stimulus or not.

3.3.2.5 Probabilistic Training Task

Previous pilots revealed that participants often have a poor understanding or representation of probabilities. The probabilistic training task started with a text explaining the concept of probabilistic feedback using the example of an 80 % – 20 % reward contingency. It was clarified that an 80 % – 20 % reward contingency means that in 20 % of the cases a correct choice will be followed by negative feedback and a negative reward (no points added), and in 20 % of the cases an incorrect choice will be followed by positive feedback and a positive reward (10 points added). It was emphasized that the 80 / 20 ratio is constant throughout the task (sampling without replacement), meaning that in every trial there was an 80 % probability to get the *true* feedback and reward and a 20 % probability to get a *misleading* feedback and reward. Participants were assured that since both correct and incorrect choices are followed by misleading feedback, the overall score is not negatively impacted by the probabilistic feedback.

The participants performed the exact same task as in the deterministic training version (*rule-based* or *stimulus-based*) with the only difference being that instead of deterministic feedback they received probabilistic feedback. Importantly, the participants knew before starting the task which stimuli were

the four valuable ones (these corresponded to the ones in the deterministic version of the task). The task only aimed to familiarize the subject with the probabilistic feedback. Participants were instructed to pay attention to the received feedback. After approximately five minutes, the task stopped.

Note that since the purpose of this task was for the participants to understand the concept of probabilistic feedback, it was performed only once. That is, each participant performed *either* a rule-based probabilistic training task or a stimulus-based training task.

3.3.3 Exposure Task

Each of the main tasks was preceded by an exposure task. The stimuli used in this task were the same as the ones used in the corresponding main task. In addition, eight *catch* stimuli were introduced which were distortions of the main stimuli (**Figure 6A**). Namely, these stimuli had features that were not present in the original stimulus set: a new color, a different orientation, or a different number of circles / stars. On each trial participants saw one stimulus

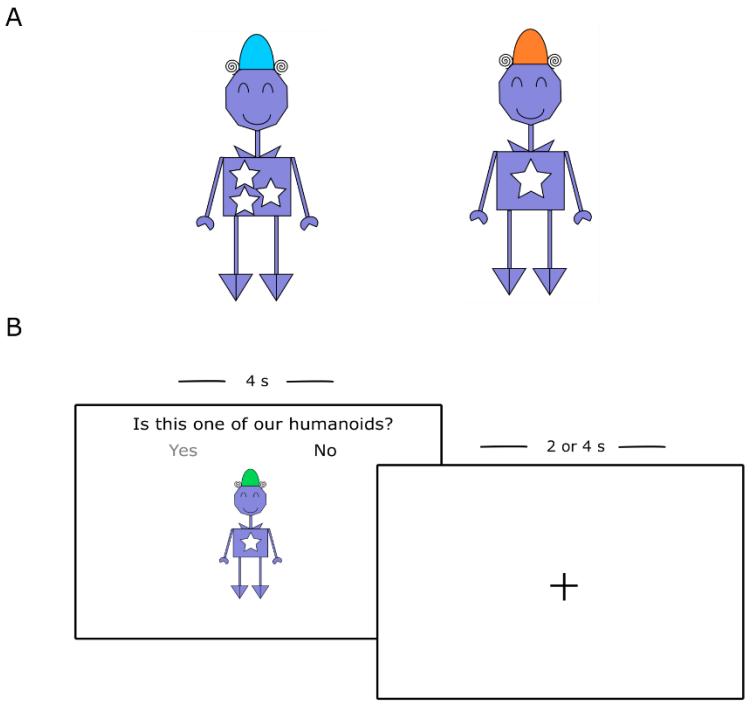


Figure 6. *Exposure task.* ***A.*** *Example catch stimuli.* ***B.*** *Example trial. Above each screen its presentation duration is displayed in seconds.*

together with the question “Is this one of our humanoids?” or “Is this one of our butterflies?” depending on the stimulus set. Participants had to choose “yes” if the stimulus presented was seen in the instruction sheets, and choose “no” if the stimulus was not previously seen. The two response options were displayed on the screen below the question. Upon selection, the chosen option was highlighted in grey. Participants had 4 seconds to decide, after which a black fixation cross would appear on the screen. The cross stayed on the screen for 2 or 4 seconds (time interval randomly chosen) and was followed by a new trial. No feedback was presented. Participants completed 40 trials containing four repetitions of the original stimuli in randomized order, intermixed with one presentation of each of the *catch* stimuli. Additionally, after every fourth trial, a null event was inserted, where a fixation cross appeared for 4 seconds instead of a stimulus.

3.3.4 Procedure

The data was collected on two separate days with the in-between testing time ranging from two days to four weeks. Prior to testing, participants were seen by a physician, who informed them about the MR safety measures and assessed whether they can safely take part in the experiment. On *Day 1* participants gave informed consent and were instructed about the type of task they were about to perform (rule-based or stimulus-based). First, they completed one of the two possible deterministic training tasks and the corresponding probabilistic training task (i.e. deterministic rule-based training task followed by a probabilistic rule-based training task). Following a short break, participants were given the corresponding instructions for the main task.

One initial eye-tracking calibration was done as soon as the participants entered the scanner to optimize head position prior to scanning. After five minutes of pre-measurements (explained in **Chapter 6.3**), a new calibration was completed and the *exposure task* started. The task lasted approximately eight minutes and participants were allowed to rest their eyes for two minutes before starting the main task. The *main task* was split into five sessions of 32 trials each lasting approximately eight minutes. Every session was preceded by an eye-tracker calibration and followed by a short break. If the participants were

performing the rule-based task, they were asked during the break if they had already found the rule and if they could describe it. No feedback was given regarding rule correctness. If the participants were performing the stimulus-based task, they were only asked to rest their eyes until the next session started. Each session lasted approximately six minutes.

Participants were only invited to *Day 2* if they successfully completed *Day 1*. On *Day 2,* subjects performed the remaining deterministic training task (e.g. if they completed the rule-based task in *Day 1*, they performed the stimulus-based task in *Day 2*) and were subsequently asked to explain the concept of probabilistic feedback. Then, the instructions for the main task were presented and clarified. The rest of the testing day continued exactly as in *Day 1*.

Participants were compensated with 30 Euros for participating in *Day 1* and 25 Euros for participating in *Day 2*. Depending on their performance, they could earn up to 5 Euros more on each testing day.

3.3.5 Data Acquisition

The software Psychophysics Toolbox Version 3 (PTB-3) running in Matlab R2014b (Neurobehavioral Systems, Inc., Berkeley, CA, USA) was used to control stimulus presentation and data acquisition. The setup contained a single PC with three external monitors: two in the control room and one in the scanner room. The main stimulus presentation screen with a resolution of 1920 x 1080 pixels was mirrored into the scanner using an MR compatible screen from the NordicNeuroLab (resolution: 3840 x 2160, pixel pitch 0.076225 (H) x 0.2247 (V), refresh rate 60 Hz). Responses were recorded using an MR compatible button box with a diamond arrangement.

3.4 Results

Examples of individual performance curves are depicted in **Figure 7**. The top and middle rows of the figure illustrate two representative subjects performing the rule-based (left) and stimulus-based (right) task. The bottom row of **Figure 7** displays two non-performers, participants who failed to perform the task (since non-performers in Day 1 were not invited to Day 2 there was no subject who failed to learn both tasks). These plots indicate that learning

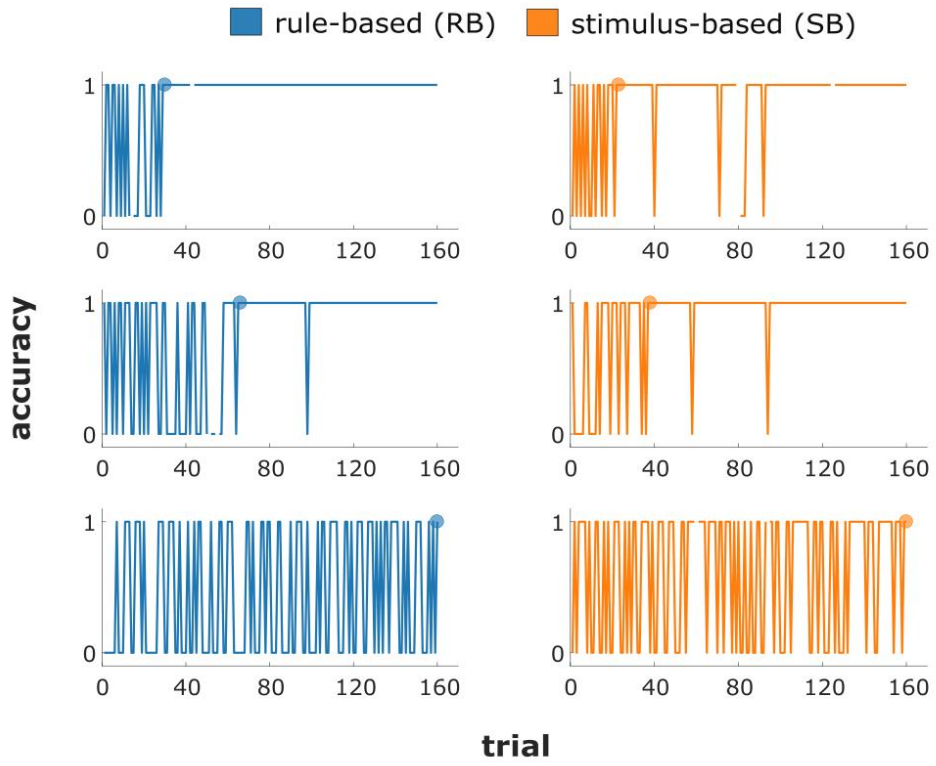


Figure 7. *Example data from individual subjects. The x-axis of each plot indicates trial number. Y-axis of each plot indicates accuracy with 0 = incorrect (subjects selected the non-valuable stimulus) and 1 = correct (subjects selected the valuable stimulus).Dots indicate learning points. Top: subject with a good performance in both tasks. Middle: subject with a medium performance in both tasks. Bottom: left, subject who failed to learn the rule-based task, right: subject who failed to learn the stimulus-based task.*

seemed to occur in an all or nothing fashion (sudden switch from mostly incorrect to perfect accuracy), particularly in the rule-based task. Noteworthy, the curves displayed do not indicate that either task was learned faster than the other.

To facilitate comparison with previous studies, the total percentage of errors in each condition was calculated. The first calculations suggested that the two conditions had a similar mean number of errors ($M_{rule\text{-}based}$ = 13.25 %, $SD_{rule\text{-}based}$ = 10 %; $M_{stimulus\text{-}based}$ = 13.95 %, $SD_{stimulus\text{-}based}$ = 6.65 %). The differences in standard deviations indicated the need for closer inspection. It was found that the rule-based mean was heavily skewed due to one participant who had a 47% error rate. This participant was a late learner and only reported the correct rule after the last scanning session. Without this participant, the overall mean error was 12.1 % and the standard deviation approached the one in the stimulus-

based task (*SD* = 7.8 %). No particularly late learner was found in the stimulus-based group (no participant exceeded 31 % error rate). From all past studies for which error percentages could be retrieved (which are in fact the studies in **Figure 4** shown in the introduction of this chapter), the present values were closest to the ones in Nosofsky et al. (1994) (in all other studies the minimum error percentage exceeded 20 % while in the current study the minimum was 4.3 %). Nevertheless, as opposed to Nosofsky et al. (1994) who found an approximately 10 % error gap between the two conditions (type II approximately 5 %, type VI approximately 15 %), this study only found a small 1.8% gap. The gap was even smaller when the slow learner was included.

The more recent studies employing these problems have opted for presenting the proportion of correct responses rather than error percentages (e.g. Lewandowsky, 2011; Mack et al., 2020). Therefore, the current section also reports proportion of correct responses. The block-wise evolution of the group proportion is displayed in **Figure 8A**. It can be observed that, although the average accuracy curves of the two tasks overlapped at the beginning, they started diverging after block 4. After this, performance in the stimulus-based task stayed below the one in rule-based task until the last block.

To further asses these observations, a mixed effects logistic regression was conducted using the *lme4* (Bates et al., 2014) and *afex* (Singmann et al., 2021) packages in R (R Core Team, 2013). A logistic regression was ran since the variable of interest, accuracy, was coded as a dichotomous variable (0 = incorrect, 1 = correct). Following Barr et al. (2013)'s advice to include all possible sources of variation in the model, the largest theoretically reasonable structure was fitted. The predictors of interest trial (all 160 trials, scaled with the R function *scale*), task type (0 = stimulus-based, 1 = rule-based) and task order (0 = started with the stimulus-based task in Day 1, 1 = started with the rule-based task Day 1) were treated as fixed effects. The stimulus type (0 = butterflies, 1 = humanoids) and the by-subject trial effects were included as random effects. Thus, the resulting regression formula was y ~ trial*task*start + (1|stim_type) + (1+trial|subject). As in Wang and Ashby (2020), *p* values were determined using the Type 3 Likelihood Ratio Test as implemented in the *mixed* function in the *afex* package. This test provides *p*-values for nonzero differences

in explained variance between the full model, containing all the predictors of interest and reduced models (**Table 1**) (Singmann et al., 2021).

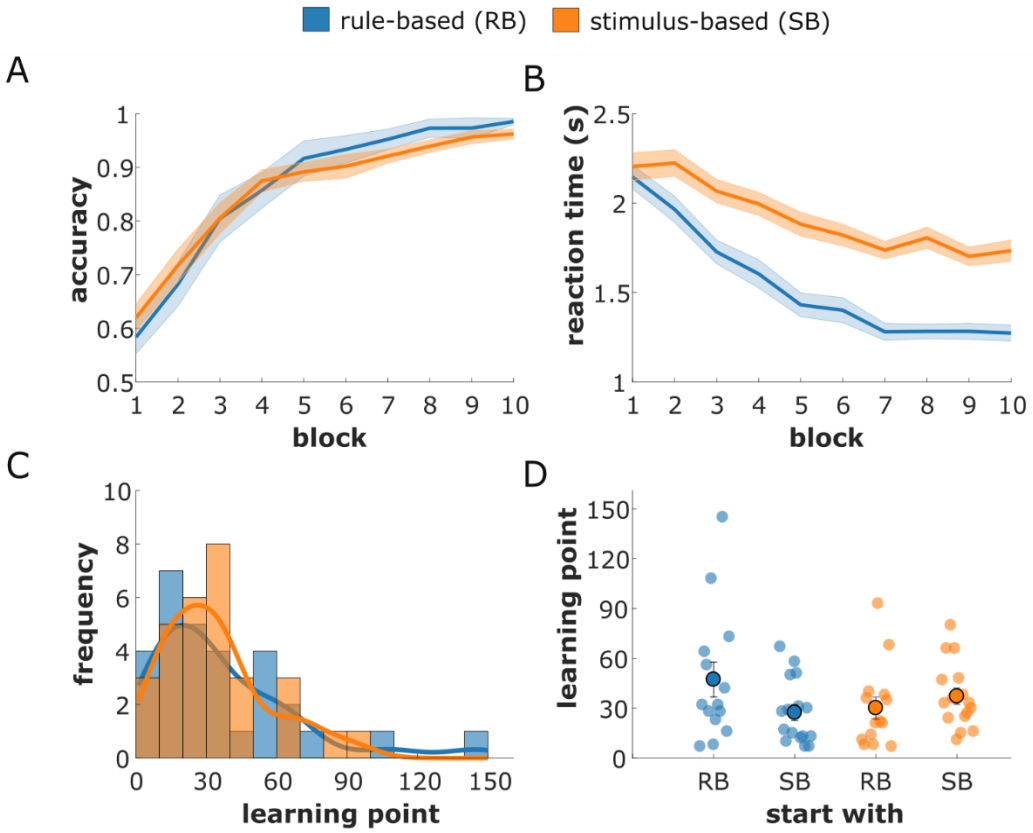


Figure 8. *Group behavioral performance (N = 30).* ***A.*** *Average accuracy in the rule-based task versus stimulus-based task. One block consists of 16 trials. On the y-axis 1 indicates perfect accuracy and 0.5 indicates chance level. Shaded areas represent standard errors.* ***B.*** *Average reaction time in the rule-based task versus stimulus-based task. One block consists of 16 trials. Shaded areas indicate standard error.* ***C.*** *Histogram of learning points in each task. Learning point was defined as the point after which the participant completed seven consecutive correct trials. Lines indicate the best fitting distributions for each task.* ***D.*** *Learning points as a function of task order (started with). Transparent circles indicate individual learning points. Filled circles indicate the respective group mean and the corresponding standard error.*

Table 1

Accuracy mixed effects logistic regression results.

Effect	df	Chi-sq	*p*
task	1	11.48	< .001
trial	1	65.09	< .001
start	1	0.58	.445
task*trial	1	26.44	< .001
task*start	1	36.26	< .001
trial*start	1	0.50	.481
task*trial*start	1	1.71	.191

Note. Type 3 Likelihood Ratio Tests (full versus reduced model).

The observations drawn from **Figure 8A** were confirmed by the regression analysis. There was a significant interaction between trial and task type, whose directionality was in line with the one depicted in the plot (*Estimate* = -0.31, *SE* = 0.11, $p < .01$). Additionally, a significant task-by-order interaction revealed that the overall task accuracy was smaller for the task performed in Day 1.

Differences between the two tasks also appeared in the reaction time (RT) data. The overall mean in the rule-based task was approximately half a second smaller than the one in the stimulus-based task ($M_{rule\text{-}based}$ = 1.54, $SE_{rule\text{-}based}$ = 0.05; $M_{stimulus\text{-}based}$ = 1.92, $SE_{stimulus\text{-}based}$ = 0.05). Moreover, **Figure 8B** indicates that although RT decreased in both tasks as a consequence of learning, the reduction in the rule-based task was more drastic (by approximately one second) than the one in the stimulus-based task. These observations were investigated further using linear mixed effects models. This time, since the dependent variable (RT) was a continuous variable following a gamma distribution, the analysis was performed using a gamma link function. The full model (presented in **Table 2**) was identical to the one used for the

accuracy data. As expected, adding the interaction between trial and task explained more variance that the model without the interaction term. The estimate of this interaction confirmed that RT reduction was less strong in the stimulus-based task than in the rule-based task (*Estimate* = 0.09, *SE* = 0.01, *p* <.001). In addition, participants that started with the stimulus-based task had a lower average RT in the rule-based task than when the rule-based task was performed first (*Estimate* = 0.04, *SE* = 0.01, *p* < .01).

Table 2

Reaction time mixed effects model results.

Effect	df	Chi-sq	*p*
trial	1	705.40	<.001
task	1	45.37	<.001
start	1	0.13	.714
task*trial	1	126.25	<.001
task*start	1	8.10	.004
trial*start	1	0.79	.375
task*trial*start	1	2.68	.102

Note. Type 3 Likelihood Ratio Tests (full versus reduced model).

It was hypothesized that the rule-based task would be learned faster than the stimulus-based task. This hypothesis was investigated by looking at learning point differences between the two tasks. Various approaches were tested to correctly identify these learning points such as a certain number of consecutive correct trials, filter and smoothing algorithms (Smith et al., 2004). The most reliable approach proved to be defining the learning point as the point in time (trial) after which seven consecutive correct choices were made (ignoring the trials in which participants failed to respond). The histograms of each task's learning points showed a considerable overlap (**Figure 8C)**. This observation

was confirmed by a negative binomial generalized linear model, justified by the learning points' distribution being best approximated by a negative binomial distribution. This model tested the effect of task type, task order and their interaction on the mean learning points. Indeed, there was no effect of task type on the learning points, indicating that the two tasks were learned at the same rate. The linear model also indicated a significant effect of task order, with participants learning faster in either rule-based or stimulus-based task when they started with the opposite task. However, this effect was more pronounced in the rule-based task and minimal in the stimulus-based task (**Figure 8D).**

3.5 Discussion

This chapter assessed Shepard et al. (1961)'s type II and type VI problems (referred to as the rule-based and stimulus-based task) in a novel choice context followed by probabilistic feedback. Behavioral analyses revealed both similarities and differences between the two category learning problems. Strong differences were observed with respect to accuracy and RT. The participants performed consistently worse in the stimulus-based task than in the rule-based task. The total number of errors participants made was smaller than in previous replication studies and matched (partially) only one previous study (Nosofsky et al., 1994). Unexpectedly, the two tasks also differed with respect to RT. The RT data showed a one second gap between the two tasks, which was caused mainly by the small post-learning reduction in reaction time in the stimulus-based task. Contrary to all previous studies, the two problems were similar with respect to speed of learning. This speed was influenced by task order, albeit the effect was mostly present in the rule-based task.

Since the type VI problem has been repeatedly found to be more difficult than the type II problem, the overall accuracy differences (although smaller than in previous studies) were not surprising. On the other hand, the reaction time discrepancies call for discussion. It is argued that these discrepancies could originate either from distinct processing times or from distinct decision times. As far as processing time is concerned, the drastic drop in reaction time in the rule-based task could illustrate that once the task was learned (correct rule was found) participants could have saved processing time by switching from

processing three dimensions to only two dimensions. This was not the case in the stimulus-based task, in which irrespective of when learning occurred, all three dimensions had to be processed in order to correctly identify the valuable stimuli. Regarding decision time, the abrupt decrease in rule-based reaction time could be a result of straightforward, time-effective rule application in determining category membership. However, in the stimulus-based task the lack of substantial reduction could be reflective of a time-costly decision strategy. Assuming a similarity-based strategy, that is, that participants were comparing the two stimuli on the screen with all stimuli stored in memory, the decision of which stimulus was the valuable one, could only be made after all similarities to previously stored valuable stimuli had been computed. Clearly, these calculations take much more computing time than simple rule application (further details of this similarity-based mechanism can be found in **Chapter 5.2**).

Nevertheless, a definite conclusion cannot be drawn without employing an experimental paradigm tailored to separate processing and decision times (i.e. Stanford et al., 2010). The categorization literature has not yet employed this type of paradigm in stimuli with clearly separable dimensions. Hence, this work could potentially inspire the development and application of these paradigms.

It is worth mentioning that the current reaction times were not in line with those found in Love (2002), neither in the overall mean values nor in the mean difference between the two tasks. It can be speculated that these differences were not simply due to paradigm change, and were instead a marker of adequate instructions. In Love (2002) instructions only specified that the category label (which was in this case a fourth stimulus feature) depended on the other three dimensions. Thus, the fact that almost equal mean RTs were found ($M_{type\ II}$ = 1.68, $M_{type\ VI}$ = 1.69) could suggest, that unlike in the current study, participants used the same strategies in the two problem types.

One significant aspect of discussion concerns the lack of difference in learning points which directly refuted the main hypothesis that the rule-based task would be learned faster than the stimulus-based task. This puzzling finding could be attributed to three of the paradigm's novel features: instruction type, display type or feedback type.

First, the idea that the instruction type impacts learning point distribution in Shepard's problem arises from the work of Kurtz et al. (2013). The authors studied the role of instruction type on the initial ordering focusing on type II and type IV problems. To reiterate, in type II problems, items can be categorized by using a disjunctive rule, whereas in type IV problems a rule-plus exception strategy is needed. It was found that the type II learning advantage was present when the participants were explicitly asked to solve the task by applying a rule, but disappeared in the absence of rule-like instructions. Nevertheless, the effect of instructions on type IV problems was not assessed. It is plausible that if the participants had been instructed about the rule-plus- exception strategy, they would have learned faster than without instructions, perhaps even as fast as in the type II problem. In a similar vein, it is also plausible that in the present study, the clear instructions to solve the stimulus-based task using a memorization strategy speeded up learning to the extent that the performance in this task "caught up" with the one in the rule-based task. This is an exciting avenue that is yet to be taken. Future work could aim to investigate the effect of instructions on all Shepard's problems. A deeper understanding of how instructions alter the initial ordering could not only advance the current understanding of these complex problems but also fine tune current category learning models (perhaps by adding an instruction type parameter).

Second, the absence of a difference between learning points could be attributed to the display type, i.e. presenting two stimuli on the screen instead of one. One could argue that the pair display could have been particularly helpful for the stimulus-based task. Assuming a similarity-based strategy, the two-pair display could have reduced the number of stimuli to be retrieved from memory by at least a factor of one (retrieving 6 stimuli instead of 7) [1] and could have led to faster learning. Even if no retrieval per se occurred, the second stimulus could have sped up learning simply by easing stimulus recognition (i.e. recognizing one of the two stimuli as being valuable or not valuable). Nonetheless, the pair

[1] Recent theories proposed that this retrieving mechanism would be too cognitively in real-life and that instead of retrieving all exemplars at once, participants either retrieve them sequentially or only retrieve the most recent ones. Nonetheless, both scenarios benefit from the presentation of two stimuli on the screen.

display should have also facilitated learning in the rule-based task by allowing participants to either exclude potential rules faster or to simultaneously test multiple rules. Thus, with learning in both tasks being potentially accelerated by the pair presentation, the display type alone could not have been the sole reason for the two tasks being learned at the same time.

From a different angle, the paired presentation could have had the opposite effect, that of hindering learning. Given that the mean accuracy in the stimulus-based task stayed below one throughout the whole task, the concern could be raised that the reason why the tasks were learned equally fast is that the stimulus-based task was not in fact fully learned. Instead of learning each stimulus and its value, the participants could have instead just memorized certain pairs of stimuli. Although rather unlikely taking into account the randomized order of pair presentation and that a stimulus appeared in more than one pair, the individual and group pair-wise accuracy were inspected to exclude this potential confound. Both individual and group data (shown in **Figure 9**) indicated that in both tasks all pairs were successfully learned, ruling out incomplete learning as an explanation for the similar distributions of the learning points.

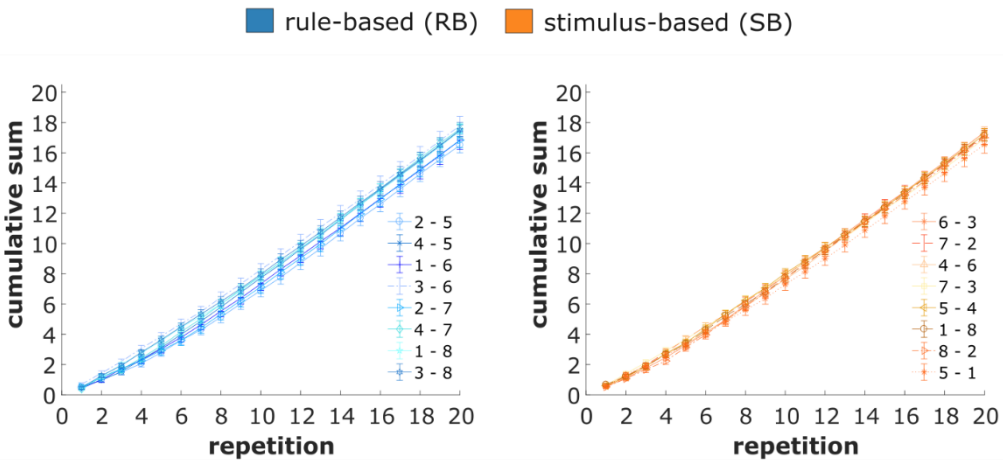


Figure 9. *Pair-wise group accuracy (N = 30). X-axis indicates the number of times a specific pair has been presented or repeated. Y-axis indicates the cumulative sum of the accuracy (a cumulative sum of 4 at repetition number 6 indicates that the participant responded correctly 4 out of 6 times when presented with that pair). Note: stimulus coding differs between the tasks, stimulus 2 in the rule-based task is different from stimulus 2 in the stimulus-based task. Error bars indicate standard error.*

Third, unlike most previous categorization tasks, the present tasks used probabilistic feedback, which could have altered the classic type II / rule-based advantage. Previous pilot tests showed that introducing probabilistic feedback slowed down learning in the rule-based task. It was assumed that since the stimulus-based task resembles a reinforcement learning task (in which learning is slowed down by the probabilistic reward scheme), introducing probabilistic feedback would have had the same effect. However, this assumption was not explicitly tested and might not have been correct. It could be the case that since the participants were aware that they cannot fully trust the received feedback, they compensated for the lack of "trustworthy" feedback by being more attentive during this task than during a standard deterministic task. This compensation could have resulted in faster learning. To rule out this possibility, the learning points in the main stimulus-based task were compared with those in the deterministic stimulus-based training task.

To reiterate, the deterministic stimulus-based training task had the same structure as the main task with the important difference that the feedback was 100 % contingent to the choice made (a correct choice would get positive feedback). The data displayed in **Figure 10** revealed that participants learned

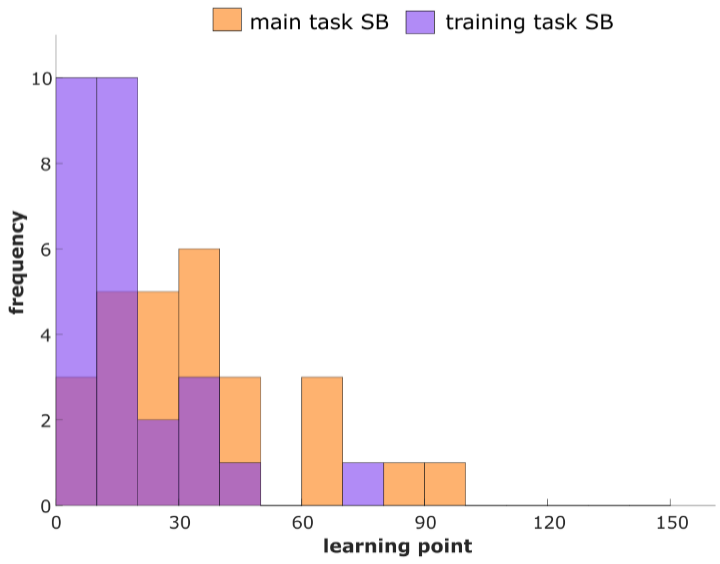


Figure 10. *Learning points distributions (N = 27). In orange, the original learning point distribution of the main stimulus-based (SB) task (main task SB). In purple, the learning point distribution of the deterministic stimulus-based training task (training SB). Learning points were defined as the point after which the subject made seven consecutive correct trials. Three subjects are missing from the original sample of 30 due to equipment malfunction.*

much faster in the training task than in the main task, thus confirming that the probabilistic feedback had the initially assumed effect. While one could argue that faster learning in the training task could be attributed to the simpler stimuli, the large difference in learning points speaks for the feedback as the main driving factor. Thus, the probabilistic feedback is an unlikely contributor to the absence of a difference in speed of acquisition.

In conclusion, the current paradigm managed to offer novel behavioral insights into rule-based and stimulus-based categorization. By taking a new approach in terms of instruction type and stimulus presentation, these tasks challenged the traditional type II and type VI problems and the established type II learning advantage. The behavioral analyses unraveled the explanatory potential of reaction times, by showing a striking difference in the development of participants' speed due to learning across the two tasks. All in all, the rule-based and stimulus-based task highlight yet again the richness of Shepard's problems and that despite six decades of research, there are many behavioral subtleties left to explore.

4. Dissociable Attentional Signatures

4.1 Introduction and Hypotheses

Categorization and selective attention are closely intermixed. Without being able to direct attention to the dimensions predictive of a certain category label, categorizers would be extremely inefficient or fail altogether (McColeman et al., 2014). Consequently, most category learning models rely heavily on selective attention mechanisms. Some of the most successful models in explaining Shepard's problems such as ALCOVE (Kruschke, 1992), RULEX (Nosofsky & Palmeri, 1998), ATRIUM (Erickson & Kruschke, 1998) and SUSTAIN (Love et al., 2004) are no exception. All these models contain attentional weights which encode the amount of attention paid to each stimulus dimension. These weights are crucial in computing associations between a stimulus and a category label. The models do however make different assumptions on how these weights are calculated and updated throughout the task. For example, the ALCOVE model proposes that attention is first allocated to all stimuli dimensions (equal attentional weights), and as the model learns the right category membership, attention shifts to the most diagnostic dimension in an error-driven fashion (the relevant dimension gets higher weights, Kruschke, 1992). RULEX on the other hand, advocates for the opposite mechanism, namely that attention is first paid to only one dimension and is progressively distributed to more dimensions if needed (Kim & Rehder, 2011). ATRIUM takes a different approach and conceptualizes attention as a mechanism that controls the interaction between a rule-learning system and an exception storing system (Erickson & Kruschke, 1998). SUSTAIN no longer separates these two systems, and instead stores both rule-learned and exception-learned items, in compartments called "clusters", whose formation is controlled by attentional weights (Love et al., 2004; Mack et al., 2018).

Understanding selective attention is essential for understanding the categorization process. It is evident that this understanding cannot be complete by using only model-based attentional estimates. Therefore, model-based estimations of attentional mechanisms should be complemented by empirical measurements of attentional processes. Selective attention in categorization

can be assessed through the use of experimental paradigms cleverly constructed to elicit different attentional mechanisms (e.g. Best et al., 2013; Carvalho & Goldstone, 2017; Deng & Sloutsky, 2016) or using indirect measures of attention (e.g. switch frequency, Matsuka & Corter, 2008). However, more direct investigations of attention can be accomplished by using eye-tracking methods. Eye-tracking has become an established measure of *overt* attention which is defined as attention-triggered eye-movements to a spatial location (Liversedge & Findlay, 2000). Although several categorization studies have successfully employed eye-tracking (Hoffman & Rehder, 2010; Kim & Rehder, 2011; Vigo et al., 2013; Carvalho & Goldstone, 2017; Zaki & Salmi, 2019), its application to the Shepard's problems has been scarce. In fact, up until now only two eye-tracking studies have investigated problems that resemble the original six problems (Watson & Blair, 2008; Blair et al., 2009) and only one study has systematically investigated attention in a subset of the original problems (Rehder and Hoffman, 2005). These studies are briefly summarized below. The study by Watson and Blair (2008) is not revised since the same data as in Blair et al. (2009) was used (with the focus on attention mechanisms during the feedback).

Blair et al. (2009) used a paradigm in which eight stimuli with binary dimensions could be categorized in four possible categories (no stimulus could belong to multiple categories). The resemblance to Shepard et al. (1961) comes from the way in which these categories were constructed: one of them could be found by using a uni-dimensional rule (akin to the type I problem), one by using a two-dimension rule (akin to the type II problem) and two by memorization (akin to the type VI problem). The authors found that participants fixated longer on the features relevant to the one- and the two-dimension rules. With respect to the memorization category, participants spent an equal amount of time fixating on two dimensions and a slightly longer time fixating on the third dimension. Moreover, it was found that the sequence of fixations within a trial mirrors the dimensions' relevance, such that relevant dimensions are fixated on first. It has to be highlighted that the authors restricted the analyses to the period after learning.

The Rehder and Hoffman (2005) paradigm was explained in detailed in **Chapter 2.1**. In short, the authors recorded eye-tracking data of standard type I, II, IV and VI problems. As in the original study, in each problem there were only two possible categories. Noteworthy, unlike the study by Blair et al. (2009), here the analyses also included the fixations *prior* to learning. It was found that in both type II and type VI problems, participants started by fixating all three dimensions. As learning progressed, participants in the type II problem switched from fixating all three dimensions to fixating only two dimensions. Unlike Blair et al. (2009), there were no post-learning fixations on a third dimension. In type VI problems, participants fixated on all three dimensions throughout the task, and no change occurred with learning. These findings were extremely important since it was the first time it was empirically demonstrated that participants learn to attend selectively to dimensions diagnostic for the problem at hand.

Given that all models used to explain the Shepard et al. (1961) problems have made powerful assumptions about attentional mechanisms, the scarcity of accompanying eye-tracking studies is surprising. Thus, this work aimed to contribute to the literature by using eye-tracking measurements of the current adaptation of type II and type VI problems. Aside from filling in a literature gap, these measurements are particularly valuable for the current study since two distinct categorization types are compared, rule-based and stimulus-based. To facilitate comparison with Rehder and Hoffman (2005), these differences will be investigated with respect to fixation *counts* rather than fixation *durations*. Given that this was the first two-stimulus display adaptation of the problems, the first step was to get a general understanding of fixation patterns in this format. As in the previous studies, the main goal was to assess whether attention was allocated optimally across the two tasks. This question was particularly important, since unlike previous studies, here clear instructions on the optimal strategies (rule application versus memorization) were provided.

Since the two problem types employ two behaviorally distinct strategies, it was expected that they would also elicit distinct attentional strategies. Due to the current work being more similar to Rehder and Hoffman (2005) than to Blair et al. (2009), the two hypotheses formulated was based on the former such that:

H1: It was expected that in the rule-based task, after learning, participants will fixate only on the rule-forming dimensions.

H2: In the stimulus-based task, it was expected that participants would fixate all three dimensions equally throughout the task.

4.2 Materials and Methods

4.2.1 Participants

Due to technical failures of the eye-tracking equipment, data from only 22 participants could be used for the rule-based task analyses (10 males, 12 females, M_{age} = 25.54, SD_{age} = 2.79) and data from only 17 participants could be used for the stimulus-based task analyses (8 males, 9 females, M_{age} = 25.53, SD_{age} = 2.62).

4.2.2 Acquisition parameters

Eye movements were recorded under constant lighting conditions from the right eye using the Eyelink Eye Tracking system (EyeLink 1000, SR Research, Ottawa, ON, Canada). The recording was done at a sampling rate of 500 Hz with a spatial resolution of 0.01 and a spatial accuracy of 0.5. In any given trial, the size of the two stimuli (both butterflies and humanoids) was 7.94 cm x 25.2 cm, which corresponded to 3.77 x 11.97 degrees of visual angle at a distance of 120 cm. The stimuli did not change in size during the feedback period. The size of the reward coin was 3.97 cm x 2.64 cm corresponding to 1.53 x 1.26 degrees of visual angle.

4.2.3 Preprocessing

The unprocessed EyeLink files were imported and converted into usable matrices using the package “eyelinker” from R (Barthelme, 2019). For both left and right stimuli, three equally sized areas of interest (AOIs) were defined (7.93 cm x 6.61 cm), each containing one of the varying dimensions (color, number or orientation) (**Figure 11**).

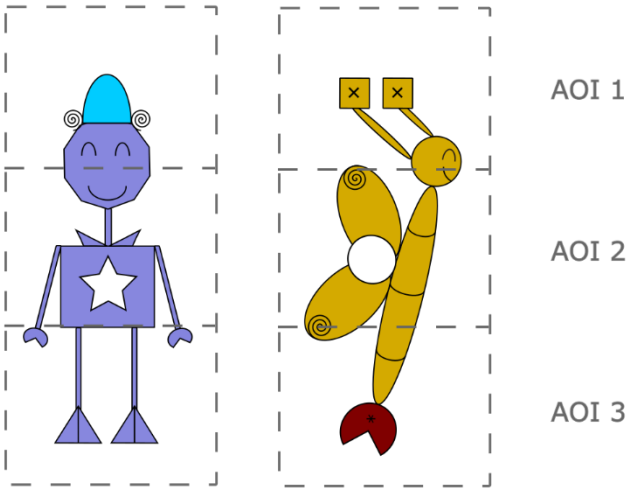


Figure 11. *Area of interest definition. Dashed grey lines mark the three areas of interest (AOI) for the humanoid stimuli (left) and for the butterfly stimuli (right). All AOIs have equally-sized areas.*

Raw data inspection indicated that the eye-tracking data suffered from considerable drifts, that is, all data points seemed to have been displaced by a certain factor from their true location (**Figure 12** left). These drifts could not be attributed to poor calibration because they also occurred in subjects with perfect calibration. To correct for this artifact, the data was feedback centered. First, the cluster of fixations corresponding to the feedback coin was identified. Second, the center of this cluster was calculated by taking the mean of all the points included. Third, the displacement factor was computed by calculating the distance between the center of this cluster and the actual center of the screen (X coordinate: 969, Y coordinate: 540). Lastly, all points were shifted by this displacement factor. Visual inspection of individual data revealed that this procedure restored the data points to their correct position and thus considerably improved the quality of the data (**Figure 12** right).

Fixations with a duration shorter than 100 ms and longer than 1300 ms were excluded from subsequent analyses (van Renswoude et al., 2018). Given that this project focused on disentangling rule-based and stimulus-based strategies, all analyses were conducted on a time window starting with the two-stimulus presentation and ending with a button press (when the participant selected one of the two figures). The choice of this time window was further motivated by the fact that on average participants responded after 1.52 seconds

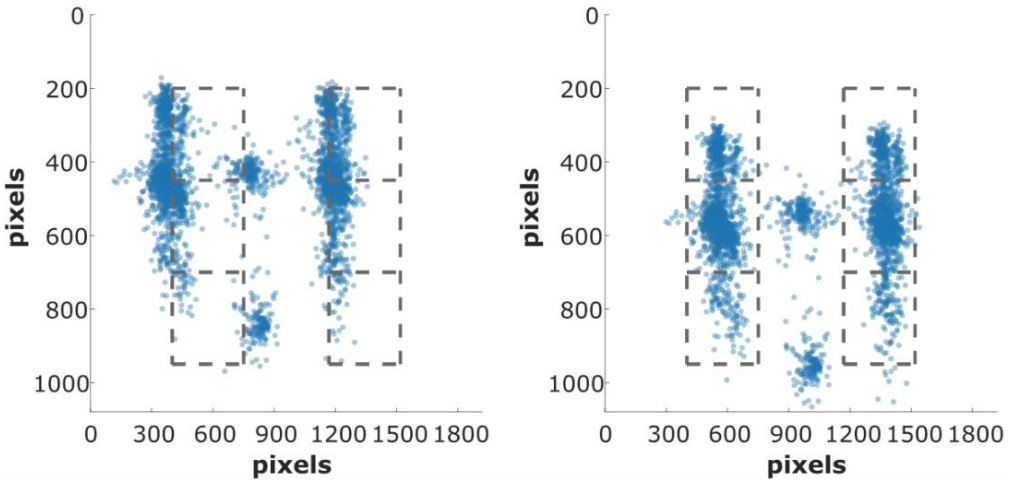


Figure 12. *Drift correction. Example data from one participant before shifting (left) and after shifting (right). Dashed rectangles indicate the stimuli location and their corresponding areas of interest as defined in* ***Figure 11****.*

in the rule-based task and 1.91 seconds in the stimulus-based task. Therefore, including the whole 4-second time window of stimuli presentation would have led to including 2 seconds of data points that were quite likely reflective of exploration or fatigue and not of the strategies of interest. Trials in which participants failed to respond were discarded.

4.3 Results

Given the novelty of these two tasks, the first analyses focused on obtaining a general overview of the fixation pattern. Here, the fixation differences between the two tasks were investigated using a generalized linear mixed effects regression with a Poisson distribution. A Poission distribution was chosen since the dependent variable, the number of fixations until choice, followed a Poisson distribution. The fixed effects were task (1 = rule-based task, 0 = stimulus-based task), stimulus (1 = humanoids, 0 = butterflies) and AOI (three levels: AOI 1, AOI 2 and AOI 3, definition based on **Figure 11**). The random effects structure included subject and by trial-subject effects (Barr et al., 2013). All in all, this resulted in the following regression formula: y ~ task*stimulus*AOI + (1 + trial|subject). This model as well as all next models in this chapter were fitted in R using the *lme4* package (Bates et al., 2014).

The results revealed a significant main effect of task, such that overall, participants needed less fixations to reach a decision in the rule-based task than in the stimulus-based task (*Estimate* = -0.48, *SE* = 0.04, *z* = -12.24, *p* < .001). Follow-up analyses showed that this difference was mainly due to participants in the stimulus-based butterfly condition needing significantly more fixations than those in the stimulus-based humanoid condition (**Figure 13A**) (*Estimate* = 2.44, *SE* = 0.17, *z* = 14.21, *p* < .001).

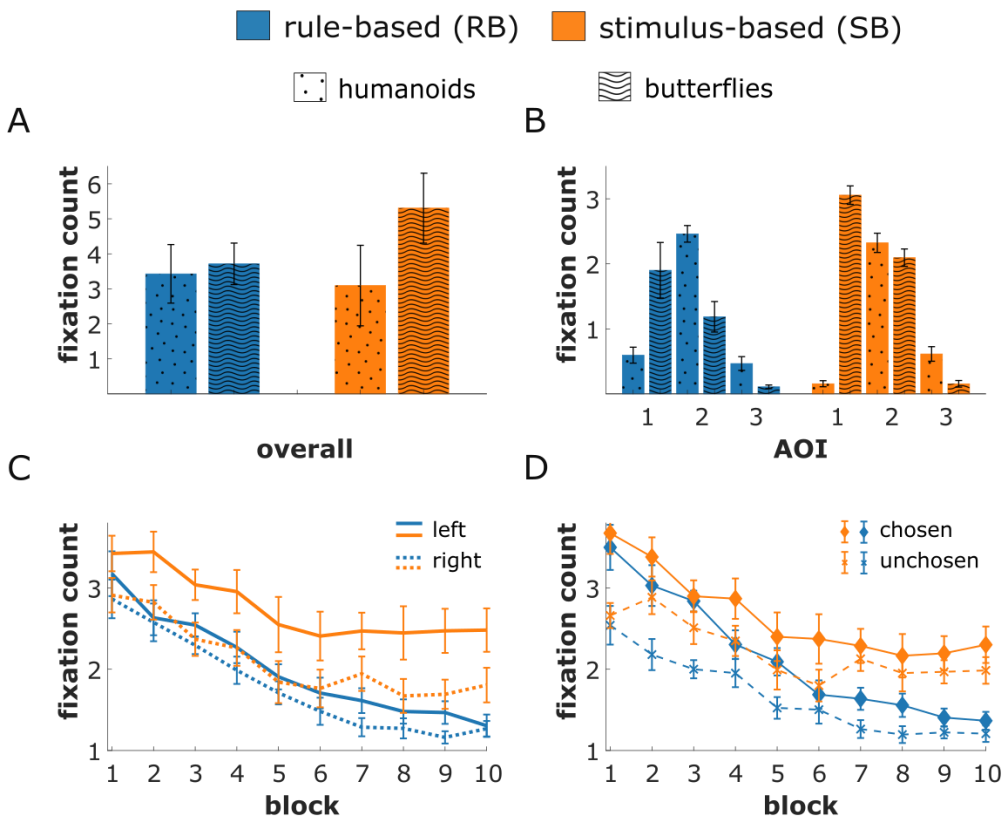


Figure 13. *Descriptive eye-tracking results (N_{RB} = 22, N_{SB} = 17).* ***A.*** *Mean fixation count in the rule-based and stimulus-based task split by stimulus type.* ***B.*** *Mean fixation count on each of the three areas of interest (AOI), split by stimulus type.* $N_{RB}^{humanoids}$ *= 16,* $N_{SB}^{humanoids}$ *= 6,* $N_{RB}^{butterflies}$ *= 6,* $N_{SB}^{butterflies}$ *= 11.* ***C.*** *Curves display block-wise evolution of the left bias between the two tasks.* ***D.*** *Curves represent block-wise differences in fixations on the chosen versus unchosen stimuli. Each block consisted of 16 trials. Error bars indicate standard errors.*

Next, the distribution of fixations across the three areas of interest was examined. **Figure 13B** indicates an AOI 2 fixation bias (center of the stimuli) which was confirmed by significant pairwise comparisons across dimensions (AOI 1 – AOI 2 with *Estimate* = -0.82, *SE* = 0.02 and *p* < .001; AOI 2 – AOI 3 with

Estimate = 1.99, *SE* = 0.37, $p < .001$). The overall fixation pattern differed significantly between the two stimuli. Regardless of task type, participants allocated almost double the amount of fixations to the AOI 2 of the humanoids (corresponding to the belly) than to either of the two external features. This was not the case for the butterflies conditions in which AOI 1 (corresponding to the antennas) received most fixations followed closely by the AOI 2 (the wings region). Due to the different instructions and strategies used in the two tasks, the separate AOIs by-task interactions were not followed up in this analysis but were the scope of a separate analysis on dimension relevance (below).

The second analysis focused on the general trial-wise evolution of the attentional allocation. As above, this analysis was also a generalized linear mixed effects model with a Poisson distribution with the number of fixations until choice as dependent variable. This time, the fixed effects were the variables: task (1 = rule-based, 0 = stimulus-based), trial (all 160 trials, scaled), left (1 = fixation on the left side of the screen, 0 = fixation on the right side of the screen) and chosen (1 = fixation on the chosen stimulus, 0 = fixation on the unchosen stimulus). The random effects structure was identical to the one in the previous analysis. Thus, the final regression equation was: y ~ task*trial*left*chosen + (1 + trial|subject).

Figure 13C shows a general reduction of fixation count with task progression but this reduction affected more the rule-based task than the stimulus-based task (significant task-by-trial interaction, *Estimate* = -0.05, *SE* = 0.03, $z = 1.99$, $p < .05$). Participants also exhibited a left bias (*Estimate* = -0.16, *SE* = 0.03, $z = -5.46$, $p < .001$) which was was more pronounced in the stimulus-based task (significant task-by-left interaction, *Estimate* = 0.15, *SE* = 0.56, $z = 2.67$, $p < .01$).

When looking at attention distribution in relation to choice, participants fixated significantly more often on the chosen stimulus rather than on the unchosen stimulus in both tasks. This gap was stronger in the rule-based task than in the stimulus based task (significant task-by-chosen interaction, *Estimate* = 0.25, *SE* = 0.04, $z = 6.65$, $p < .001$). **Figure 13D** depicts that in the stimulus-based task, the fixation count on the chosen and unchosen stimuli, as well as the difference between the two, stabilized after block 6. In the rule-based task on the other hand, both the fixation count and the difference continously

decreased to such an extent that in the last block there was no longer a difference between chosen and unchosen stimuli.

Once the general fixation pattern was described, the next analyses assessed whether the fixation allocation to the stimuli' dimensions were indeed reflective of the two different strategies used in the two tasks.

To reiterate, it was expected that in the rule-based task participants would learn to allocate attention to the two dimensions forming the disjuctive XOR rule (e.g. the color and number dimensions when the correct rule was blue hat and one star or green hat and two stars) and to disregard the dimension which is not part for the rule (e.g. the orientation dimension). To test this hypothesis, the data had to be recoded according to dimension relevance. As described in **Chapter 2.3.4,** each participant was assigned a task version that could be solved with one of three possible rules. The valuable stimuli could be correctly identified either by a combination of color and number (N = 6), color and orientation (N = 8) or number and orientation (N = 9). Since the main purpose of the task was to test differences in attentional allocation between the relevant and irrelevant dimensions, all fixations were re-grouped based on their role within the rule-forming dimensions. Based on its location, each fixation was assigned to one of three possible groups: *relevant 1* or *relevant 2* if it was located on one of the two dimensions forming the rule, or *irrelevant* if it was on the third dimension which was not relevant to the rule. For example, if the correct disjunctive rule was a combination of color and orientation (i.e. humanoids with blue hats and upwards shoes are valuable), the fixations on AOI 1 (the AOI corresponding to the color dimension for humanoids) were assigned to the *relevant 1* group. The fixations on AOI 3 (the AOI corresponding to the orientation dimension for humanoids) were assigned to the *relevant 2* group, and the fixations on AOI 2 (corresponding to the number dimension for humanoids) to the *irrelevant* group (**Figure 11**). Note that this was done for both stimuli presented on the screen on each trial. Fixations outside of these AOIs were not included in the analysis. The recoded data was inspected using a generalized linear mixed effects model with a Poisson distribution. As in the analyses above, the dependent variable was the number of fixations. The fixed effects of interest were trial (all 160, scaled), relevance (0 = relevant 1, 1 =

relevant 2, 2 = irrelevant) while the random effects included subject and by-trial subject effects. Thus, the final regression formula was y ~ trial*relevance+(1+trial|subject).

The analysis revealed an overall significant difference between all three levels of relevance, with the strongest discrepancy being seen between the *relevant 1* and *irrelevant* dimensions (*Estimate relevant 1 – relevant 2* = -.25, Estimate relevant 1 – irrelevant = -0.93, *Estimate relevant 2 – irrelevant* = -.67, $p < .001$ for all comparisons). **Figure 14A** displays the evolution of fixations on the relevant and irrelevant dimensions. It can be seen that while fixation count on all dimensions decreased with task progression, the irrelevant dimension shows the steepest trial-wise decrease. This was confirmed by the lack of a significant trial-by-relevance interaction for the relevance dummy coded group (reference group: relevant dimension 1) and a significant trial-by-relevance interaction for the irrelevant dummy coded group (reference group relevant dimension 1, *Estimate* = -0.41, *SE* = 0.03, z = -16, $p < .001$). Noteworthy, while the number of fixations on the relevant dimensions seemed to stabilize around block 6, the number of fixations on the irrelevant dimension continued to drop.

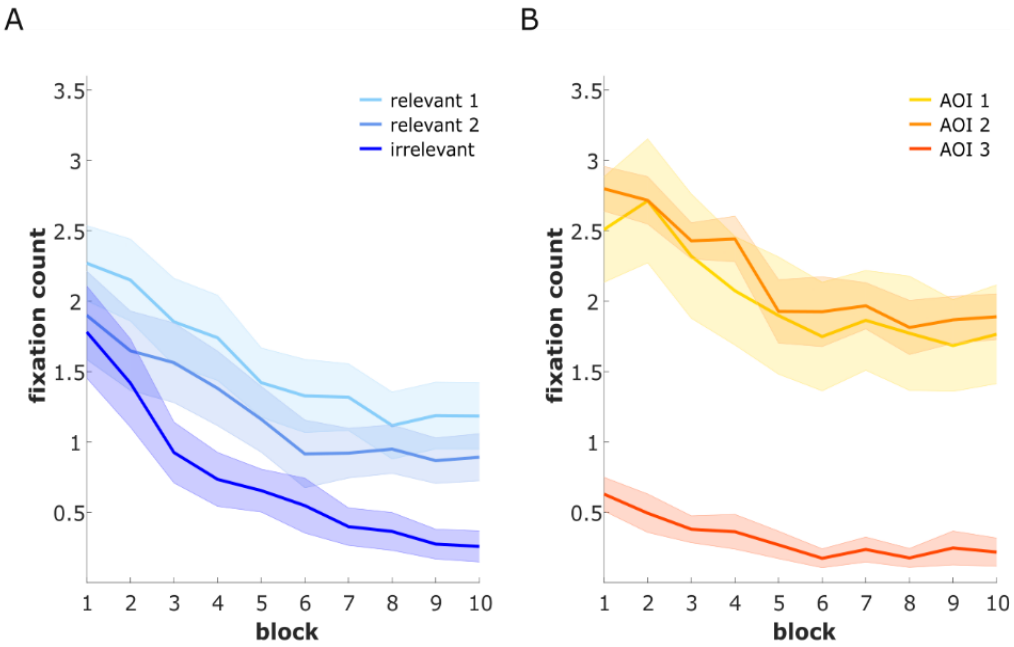


Figure 14*. Strategy-dependent attentional allocation.* ***A.*** *Rule-based task. Curves indicate block-wise evolution of attentional allocation to the relevant and irrelevant dimensions (N = 22).* ***B.*** *Stimulus-based task. Curves indicate block-wise attentional distribution to each area of interest (AOI). Each block consists of 16 trials. Shaded areas indicate standard errors.*

Concerning the stimulus-based task, it was hypothesised that since no single dimension is predictive of category membership, participants would distribute their attention evenly to all three dimensions. Therefore, since no dimension was more relevant than the other, the original fixation location coding was kept. As in the rule-based task, this hypothesis was investigated by looking at the trial-wise evolution of fixations on the three AOIs. A generalized linear mixed effects model with a Poisson distribution was applied, with fixation count as dependednt variable, trial (all 160 scaled) and AOI (3 levels, 0 = AOI 1, 1 = AOI 2, 3 = AOI 3) as fixed effects, and subject and the by-subject trial effects as random effects. The final model equation was y ~ trial*AOI + (1 + trial|subjects).

Figure 14B indicates that on average fixations were distributed evenly on AOI 1 and AOI 2, although AOI 1 had considerably more variation. The regression results confirmed this observation through a small significant difference in mean fixation count (dummy variable AOI 1 – AOI 2, Estimate = 0.06, SE = 0.02, p < .001). On the other hand, AOI 3 had a rather small mean fixation count, which differed largely from AOI 1 and AOI 2 (dummy variable AOI1 - AOI 3, *Estimate* = – 1.92, *SE* = 0.03, *z* = -49, *p* < .001).

Figure 14B also shows that fixation counts on AOI 1 and AOI 2 slightly decreased during the first five blocks and only stabilized from block 6 onwards. Nevertheless, this small block effect was not significant (non-significant trial x dummy variable AOI 1 – AOI 2 interaction). In contrast, the fixation count on AOI 3 dropped more steeply than the other two AOIs and this effect reached significance (significant trial-by-dummy variable AOI 1 – AOI 3 interaction, *Estimate* = -0.24, *SE* = 0.03, *z* = -6.41, *p* < .001).

4.4 Discussion

This chapter focused on understanding attentional allocation in rule-based and stimulus-based learning. This study was the third eye-tracking study employing Shepard et al. (1961)'s problems, and the first to address them with the atypical two-stimulus display.

The descriptive analyses suggested that in terms of attention, the stimulus-based task was more taxing than the rule-based task. This aspect was

highlighted by three findings. Firstly, the participants needed more fixations to reach a decision in the stimulus-based task than in the rule-based task. Secondly, although a left bias was present in both tasks, this bias was considerably more pronounced in the stimulus-based task. Thirdly, the fixation gap between the chosen-unchosen stimuli was only minimal in the stimulus-based task as opposed to the rule-based task, indicating that participants had to pay considerable attention to both stimuli from the beginning to the end of the task. Unfortunately, no direct comparison to previous data was possible since the study by Rehder and Hoffman (2005) did not provide mean fixation count data. Indirect connections can however be drawn to Blair et al. (2009), which found that participants fixated longer on stimuli belonging to the categories that required memorization than to the stimuli belonging to the category that required a two-dimension rule. It could be speculated that this finding also supported the claim that stimulus-based categorization has higher attentional demands than rule-based categorization. Even without these indirect connections, it can be argued that since the type VI problem has been consistenly found to be more difficult than the type II problem, it was to be expected that its increased level of difficulty would translate in higher attentional demands.

To get a better understanding of the attentional differences between the two tasks, the data was analyzed with respect to the two distinct categorization strategies. The results revealed signatures of strategy-dependent attentional allocation, albeit in different ways than hypothesized.

In the rule-based task it was expected that after learning, participants would completely disregard the irrelevant dimension and only fixate on the relevant dimensions. The findings from the relevance analyses indicated that this hypothesis was only partially confirmed. Although, as expected, most fixations were allocated to the relevant dimensions after learning, fixations on the irrelevant dimensions did not completely cease. Surprisingly, the fixation count on the irrelevant dimension continued to fluctuate long after the fixation count on the relevant dimensions had stabilized.

The finding that all three dimensions were still attended by the end of the task is in disagreement with the previous work which showed that only two

dimensions were attended by the end of the type II problem. It could be argued that this discrepancy could be solely attributed to the fact that the current study used two different stimulus sets and not the standard one stimulus set used in Rehder and Hoffman (2005). This idea is strengthened by the results showing that the humanoids and butterflies differed in their overall fixation pattern. To rule out this possibility the by-stimulus group data was investigated. Although data loss did not allow for a thorough statistical analysis of the stimulus effect (uneven small groups $N_{RB}^{butterflies}$ = 6, $N_{RB}^{humanoids}$ = 16), visual inspection confirmed that in both stimulus groups the fixation count on the irrelevant dimension fluctuated long after learning but did not reach zero.

With the stimulus confound ruled out, it follows that the presence of fixations on the irrelevant dimension together with the steady decrease were indicative of an additional attentional mechanism. One simple explanation could be that since participants had to attend to two stimuli instead of one (as in Rehder & Hoffman, 2005), the fixations on the irrelevant dimension were just explorative and motivated solely by a "getting to know" the environment mechanism. However, the underlying mechanism could be more complex and could indicate a strategy optimization attempt. Given that in the second half of the task the participants needed on average only a small number of fixations to choose the valuable stimulus, the third fixation could be a shortcut to rule-application via recognition. In certain trials participants could find it easier to make one fixation to each dimension in the hope of recognizing the stimulus as valuable or not, rather than applying a rule. Nevertheless, either due to the clear instructions to solve the task by rule application or due to the shortcut failing, the participants could be gradually giving up this avenue – hence, the continuous block-wise drop.

On the other hand, rather than a tentative optimization, attention to the third dimension could be a consequence of probabilistic feedback. As described in **Chapter 2.3**, the pseudo-randomized probabilistic feedback ensured that up until the end of the task subjects will occasionally receive negative feedback (even though they chose the correct, valuable stimulus). A recent study by Arbel et al. (2020) showed that negative feedback results in a higher change in fixation probability (whether an AOI will be fixated or not) than positive feedback. This

finding could be directly translated to the current task by attributing the third fixation to a post-negative feedback caution. In other words, even after the correct rule was found and successfully applied, the impact of the negative feedback could have been so strong that it resulted in a short-term rule doubt, and a reassessment of all dimensions. It could have also been the case that this fixation was an involuntary error caused by the negative feedback, which got corrected in the upcoming trials.

It has to be mentioned that although their paradigms were considerably different from the current one (not only in the number of categories but also in category size), Blair et al. (2009) did find that after learning participants also attended to the irrelevant dimension. Moreover, the mean duration of these fixations was far from negligible (approximately 400 ms when the fixation on the relevant dimensions had a mean of 800 and 500 ms). Due to the regularity at which they occurred in a trial sequence, the authors suggested that they were "essentially noise that was uniformly distributed in time" (Blair et al., 2009, p. 1203). This conclusion is weakened by the present findings. Although it is difficult to pinpoint the source of these fixations (especially considering that Blair et al. (2009)'s paradigm had neither two-stimulus display nor probabilistic feedback), there seems to be more to the story than noise or artifacts. It is hoped that this aspect will be researched further in the future.

As far as the stimulus-based task is concerned, the hypothesis that all three dimensions would be attended equally was also not fully confirmed. Even though participants did pay attention to all three dimensions, this attention was not distributed evenly. A comparison to the previous study was useful in understanding this aspect. At a first glance, the current results also seemed to disagree with Rehder and Hoffman (2005). The authors found that in the type VI problem all three dimensions were fixated throughout the task. **Figure 14B** depicts that in this study participants focused on two dimensions and paid little to no fixation to the third dimension. While this distinction could be introduced by the difference in display type (one versus two stimuli), the more likely explanation is that the low fixation count was a color artifact. Despite the fact that the low sample size did not allow for a thorough follow-up analysis, visual inspection of the by-stimulus group data supported this idea. This data indicated

that in the stimulus-based humanoids group, AOI 1 (corresponding to the color dimension in humanoids) had the lowest, zero approaching fixation count while in the butterflies group, AOI 3 (corresponding to the color dimension in the butterfly group) had the lowest fixation count. Since the butterfly group was almost twice as big as the humanoids group (due to data loss), the average data was pushed towards the mean of the butterfly group. Assuming a color artifact, it remains that the participants indeed allocated attention evenly to all three dimensions as in Rehder and Hoffman (2005), but the attention on the color dimension was only covert, and thus not reflected in the number of fixations.

It has to be highlighted that the differences between the current tasks and Rehder and Hoffman (2005) could be due to three important methodological differences. First, the current study only included fixations during the *choice* period, namely fixations which occurred in the time window between stimuli presentation and response. It is not clear which time window was used in the previous study. It is not specified whether the analyzed fixations occurred only during the "categorization time" (starting from stimulus presentation and ending with response) or whether all the fixations within a trial were considered, including those from the "feedback time". If the latter is true, unlike the current study, the previous findings could be reflective of not only the strategies used to categorize the stimulus, but also of those used to integrate the feedback. Previous research (Arbel et al., 2020; Blair et al., 2009) has shown that the attentional mechanisms during feedback are complex and depend on various factors, such as response correctness, which might not be as relevant to the categorization process per se.

Second, in Rehder and Hoffman (2005) participants finished the task after they completed 32 correct trials in a row. To facilitate comparison across performances, the authors "assumed that their eye movement data for the remaining blocks would have been identical to their last actual four blocks" (Rehder & Hoffman, 2005, p. 978). The authors report an average stopping point at 14.11 blocks in the type II and 22.94 blocks in the type VI problem (a block contained eight trials). The average curves presented in the paper ranged up to block 28. This implies that for an average of 14 and 6 blocks respectively, the evolution of attentional allocation was only assumed. This was not the case in

the current study where attention was monitored for the same fixed number of trials, and thus the attentional curves indicate the true mean. It could be speculated that attention could have evolved differently than assumed in the previous study, perhaps more similarly to the current tasks. This aspect would have been particularly interesting in the rule-based task in order to see if participants undergo further rule optimization.

Third, this study differed from the Rehder and Hoffman (2005) with respect to instructions. As mentioned in **Chapter 2.1**, the authors did not specify how the participants were instructed, and it can only be deduced from the discussion that no rule-like instructions were given. If this were the case, one could expect that the different instruction type led to different attentional allocation strategies. Nevertheless, the strongest effect of instruction type would have likely appeared at the beginning of each task. That is, participants in the present study would have started with another fixation pattern than those in the previous study. However, this was not the case since in both studies participants started by fixating all three dimensions. Instruction could have nonetheless impacted attention at a later stage. Without further research on effects of instruction type on categorization, it is difficult to speculate what this effect might be.

While the two tasks clearly differed in their attentional demands and patterns, some similarities did arise. Firstly, in both tasks participants distributed attention based on the instructed optimal strategy. Secondly, without the color bias in the stimulus-based task, it can be argued that in both tasks participants started by fixating all three dimensions almost equally. This argument would confirm Rehder and Hoffman (2005)'s conclusion that rather than sequentially attending to one dimension until its diagnostic value is established (hypothesis testing), participants start by fixating all dimensions. The current results strengthen their conclusion by proving (in the rule-based case) and assuming (in the stimulus-based case) that the "all dimensions start" also exists when participants are explicitly instructed on how to perform the task. Thirdly, in both tasks the fixation pattern did not stabilize after learning. Instead, despite the fact that learning was completed on average by block 4, the fixation count on the relevant dimensions (rule-based task) and on the three AOIs (stimulus-based

task) only stabilized after block 6. Interestingly, a similar phenomenon was also reported by Rehder and Hoffman (2005). The authors found that fixation on the relevant dimensions followed but not preceeded perfect accuracy. Nonetheless, in their study attentional allocation "settled" shortly after learning. One could assume that in both studies learning was immediately followed by a process of optimization, either of the current strategy or its implementation. The fact that this process took two blocks in this study rather than a few trials as in the previous study could be attributed to the two-stimulus display together with the probabilistic feedback. In other words, subjects simply had more optimization strategies to try in the current paradigm than in the typical one-stimulus deterministic setup.

Although the eye-tracking data has been in general successful at capturing attentional strategies unique to each task, there are some noteworthy limitations to the study. As previously mentioned in this chapter, the data set suffered from considerable technical problems during acquisition, resulting in the exclusion of many participants. This on its own hindered the balance of the design and impaired subgroup analyses by stimulus type (in both tasks) or by rule-type (in the rule-based task). The relatively small overall number of fixations in both tasks could have been caused by data loss. Unfortunately, since no previous study tested Shepard's problems with two stimuli on the screen and Rehder and Hoffman (2005) did not report fixation counts, one can neither confirm nor rule out this possibility. Moreover, having the dimension color in the stimulus set hindered the ability to observe fixation allocation on AOI 3 in the stimulus-based task and quite likely affected the overall fixation count in the rule-based task as well. While the color dimensions added more ecological validity to the study (in real-life when these problems are encountered the items to be categorized are almost always colored), perhaps at such an early stage of the eye-tracking literature on Shepard's problems, it would be advisable for future studies to either avoid using color as a varying dimension, or include color variations in more dimensions. The vertical format of the stimuli was also a potential caveat. Although the stimuli were more ecologically valid than the text symbols used in Rehder and Hoffman (2005) or the amoeba-like microorganisms used in Blair et al. (2009), the vertical positioning of the relevant

dimensions contaminated the data with microsaccades via the middle dimension (AOI 2). While it is hoped that most of these microsaccades were eliminated by setting the threshold for fixation at 100 ms, an AOI 2 bias was still found (**Figure 13B)**. It follows that one cannot disentangle whether this bias was a stimulus confound or an attentional strategy. It is encouraged that next studies would take a step back to triangularly-structured stimuli to avoid this potential confound.

In conclusion, despite its potential limitations, the current study was able to capture differences in attentional allocation between the rule-based and stimulus-based tasks. While attention (covert or overt) was distributed evenly across dimensions in the stimulus-based task, the rule-based task was characterized by more fixations to the relevant dimensions than to the irrelevant one. It is hoped that the findings of this study will contribute to a better understanding of type II and type VI problems and that this work was a first step in reviving eye-tracking research on the Shepard et al. (1961)'s problems.

5. Dissociable Computational Signatures

5.1 Introduction

The importance of Shepard et al. (1961)' problems and their ordering in category learning models has been paramount. Numerous models have been developed either based on the problems or with the main goal of explaining their ordering. These models have "come in many flavors" (Blair et al., 2009, p. 330), some of which are: exemplar models (e.g. GCM and extensions, Nosofsky, 1986; ALCOVE, Kruschke, 1992), cluster models (e.g. the rational model, Anderson, 1991; SUSTAIN, Love et al., 2004), connectionist models (e.g. configural cue-model and extensions, Gluck et al., 1989; DALR, Gluck et al., 1992; EXIT, Kruschke, 2001; DIVA, Kurtz, 2007), rule models (e.g. RULEX, Nosofsky & Palmeri, 1998), hybrid models (e.g. ATRIUM, Erickson & Kruschke, 1998), and software agents (e.g. CBSA, Pape & Kurtz, 2013).

It was established that one of the biggest challenges for the above-mentioned categorization models was to be able to capture the type II advantage, the ease of acquisition of type II problems with respect to the other problems, in particular with respect to type IV problems. Therefore, the most successful models were deemed those who were able to capture the nuances of this advantage: ALCOVE, DIVA, RULEX and SUSTAIN. The work by Kurtz et al. (2013) called the type II advantage into question and discovered that its evidence is far from unequivocal. In fact, the advantage occurs only under certain experimental manipulations, specifically, when rule-like instructions are given and the stimulus set contains easily verbalizable dimensions. The authors proposed that these findings cast doubt on the explanatory power of previously successful models, since these models would have difficulty in capturing an equality in ease of acquisition between the type II and type VI problems. Moreover, past models would likely fail in predicting cases in which the type IV problems would be learned faster than type II ones (type IV advantage). The concerns of Kurtz et al. (2013) have remained unaddressed until recently, when a new category learning model, Categorization Abstraction Learning (CAL, Schlegelmilch et al., 2021) solved both type II and type IV advantages by implementing an interaction mechanism between rule prediction and

memorization. The model explained each advantage by a stronger usage of either the rule-learning or the memorization mechanism.

Another previously unaddressed critique to the state-of-the art models was the distribution of selective attention. The common denominator of these models was the assumption that optimal distribution of selective attention precedes the cessation of errors. The eye-tracking study by Rehder and Hoffman (2005) provided empirical evidence that the order of these processes is in fact reversed by showing that attention was still paid to irrelevant features after optimal performance had been achieved. CAL resolved this discrepancy by assuming self-confirmatory attention learning as opposed to the standard error-driven attention used by previous models. Put simply, the model proposed that first the features most predictive to the correct classification rule are determined, and then attention is allocated to these features, thereby "self-confirming" their importance.

This work applied the newly developed CAL to the current two-stimulus display adaptation of the type II and type VI problems. Three main arguments deemed CAL the most suitable candidate for explaining performance in the newly designed categorization task. First, CAL's rule prediction and memorization mechanisms mirror the types of learning of interest in the current study, rule-based and stimulus-based categorization. Second, CAL's ability to capture a lack of type II advantage was a strong argument for employing the model since the behavioral findings (**Chapter 3**) also refuted this advantage. Third, the implementation of self-confirmatory attentional learning, as opposed to error-driven attentional allocation, aligns with the eye-tracking findings in **Chapter 4**, namely that participants still allocated fixations to the irrelevant dimensions long after the learning criterion had been reached.

Due to the novelty of the current paradigm and of CAL itself, the present work was exploratory. The main goals were to assess whether CAL could reliably capture participants' behavior in the new experimental paradigm, and whether the model could reveal new insights about the rule-based and stimulus-based task, at a behavioral as well as at a neural level.

5.2 Category Abstraction Learning

CAL is thoroughly introduced in Schlegelmilch et al. (2021). Each of its aspects is beautifully argumented from both conceptual and methodological perspectives. This section only aims to give a general understanding of the model's core features. Notably, the current description was restricted to stimuli with two discrete features belonging to two possible categories. The model can also accommodate continuous features, as well as more categories.

CAL is built on two networks: rule network and configural memory. As their names suggest, the rule network is responsible for rule-like learning, whereas the configural memory network is responsible for memory- (stimulus-) based learning. Before each network is explained in more detail, four of CAL's mechanisms warrant explanation: similarity- and dissimilarity-based generalization, rule-learning, contextual modulation and feature attention.

5.2.1 Key Concepts

In terms of learning, CAL assumes that participants learn simultaneously about present and absent categories. Learning about the present category occurs via *similarity-based generalization* (akin to GCM) and learning about the absent category occurs via *dissimilarity-based generalization* (contrasting). The following example helps to distinguish between the two. A set of objects with five possible lengths is given: 1, 2, 3, 4 or 5. Out of this set, the object with a length of 1 was classified as K. The categories of the other objects are unknown. The model learns about the category membership of the other objects by computing psychological distances between the object with a length of 1 and all remaining objects[2]. These distances are indicators of similarity and dissimilarity of the objects with respect to the object of length 1. CAL learns about objects belonging to category K (w in **Figure 15** coming from each feature) by computing similarities to the object of length 1. For example, objects with length 2 or 3 are highly similar to the object with length 1 and thus more likely to belong

[2] As in all similarity-based approaches, a distance between a stimulus and itself is also computed. This aspect is omitted from the explanation for simplification but it is included in the model.

to category K. Conversely, the model abstracts about the absent category $not\ K$ by computing dissimilarities to object 1. Thus, stimuli with length 4 or 5 which are highly dissimilar to the object of length 1 are more likely to belong to category $not\ K$. CAL assumes that the behavior of this generalization, represented by the generalization gradient , is directly correlated to *rule-learning*. For example, if the generalization from object 1 to object 5 is steep (referred to as a narrow gradient), this is indicative of a strong rule such as: objects 1 and 2 belong to category K and objects 3, 4 and 5 belong to category $not\ K$ (akin to a softmax function with a high temperature). However, if the generalization is gradual (wider gradient), the previously mentioned rule becomes weak, and there is no strong delineation between category K and category $not\ K$ (objects 2, 3, 4 and 5 cannot be separated - akin to softmax function with a low temperature). In other words, a wide gradient indicates weak rules or the absence of rule-learning. It is important to note that CAL's conceptualization of similarity and dissimilarity-based generalization as two inversely related functions of the same mechanisms is a novel approach. Previous category learning models have treated them as distinct mechanisms.

Another important aspect about *rule-learning* in CAL is that the model has a preference for simple, uni-dimensional rules. When a uni-dimensional rule fails, CAL does not correct it, as previous error-driven models do. Instead, CAL stores it unaltered and searches if the rule failure can be attributed to the context. A good example is the rule "new is better". In category learning terms this rule can be translated as: objects with the feature "new" on the *age* dimension belong to the category "better". This rule is successful for items that take on the feature "electronics" on the *type* dimension (e.g. computers) but it fails for items with the feature "alcoholic drinks" on the *type* dimension (e.g. whiskey). Thus, one learns that the rule "new items" indeed predicts the category "better" but in the context of "alcoholic drinks" its inverse has to be applied: "old is better". CAL refers to this process as *contextual modulation* and implements it through variables called contextual modulators, whose role is to adapt the rule predictions to the right context. A note needs to be made that CAL imposes the constraint that a given dimension can be either modulated or a modulator.

The contextual modulator mechanism also has implications at the attentional level, precisely on how attention is distributed to each feature, referred to as *feature attention*. That is, subjects have to keep track not only of the extent to which a certain dimension predicts categorization, but also of the context in which the prediction is not successful and needs to be reversed. The model takes this into account by computing two distinct attentional variables: one coding for attention to an object' dimensions α and one coding for attention to the context β. The values of these variables depend on how successful a dimension was in predicting the correct classification (as a rule-predictor or as a modulator) in the previous trials.

Next, each network is explained in more detail by using the example depicted in **Figure 15**. Since CAL has a preference for uni-dimensional rules, the description starts with the rule network. The configural memory network is introduced next, followed by the calculation of the overall prediction from the two networks.

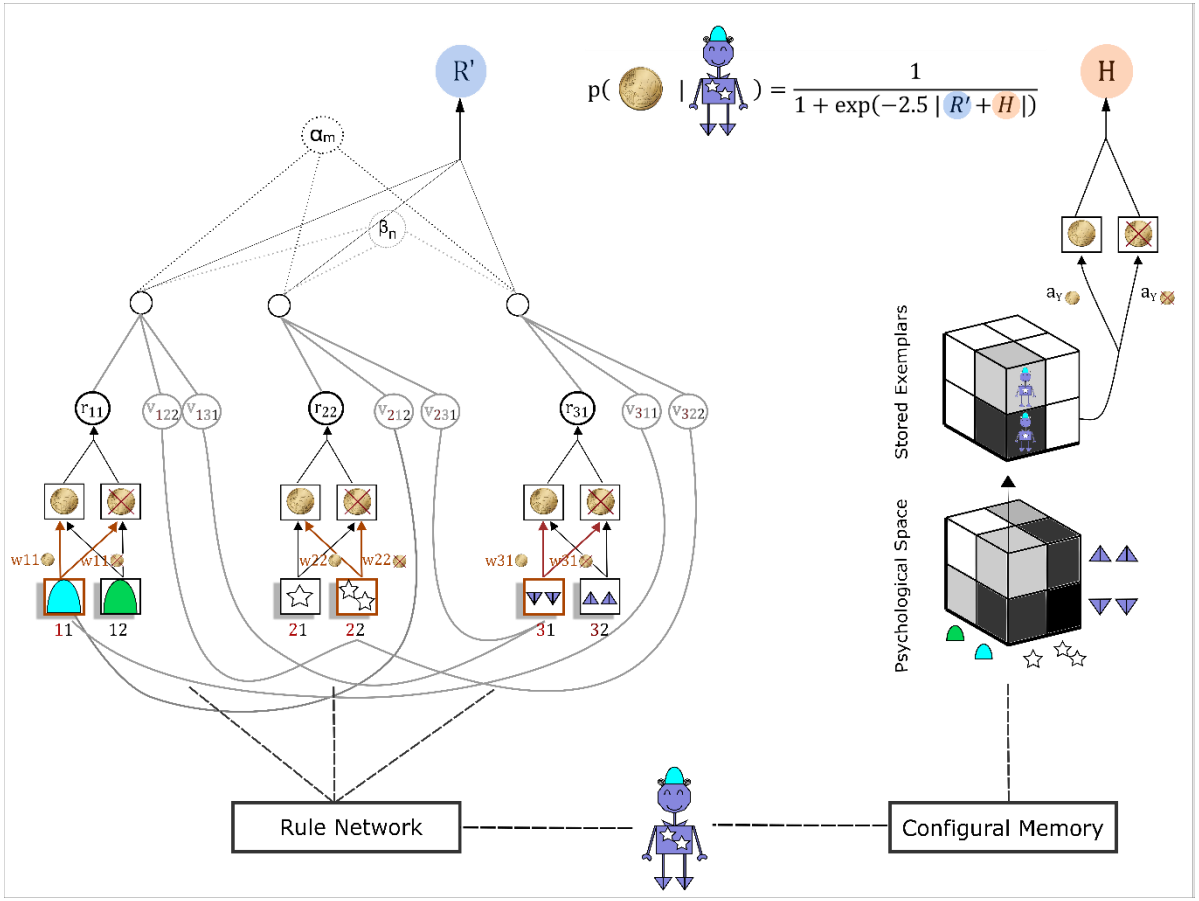


Figure 15. *Category Abstraction Learning(CAL) model. Example of how CAL processes an incoming stimulus with a blue hat, two stars on the belly and shoes pointing downwards.* ***Left:*** *rule network. In each trial, CAL learns about present features (i.e. $w11$) and abstracts about absent features (i.e. w_{12}). Evidence weights are computed for both present (valuable) and absent (not valuable) categories (i.e. w_{11} and w_{11}). For each of the current stimulus' features, a ratio r is computed by taking the logarithm of the evidence for the present category divided by the evidence for the absent category. These ratios are later weighted by how much attention is paid to the specific dimension (α_m), whether this dimension is contextually modulated (i.e.v_{122}) and how much attention was paid to the specific modulator (β_n). The overall prediction of the network R' is the sum of these weighted ratios.* ***Right:*** *configural memory. When the rule network fails, CAL switches to memorization. Category associations are computed for both present and absent categories (a_Y and a_Y) by comparing the current stimulus with all stimuli previously stored in memory. These associations are used to compute the overall prediction H. Throughout the figure, the coin indicates the valuable category, the crossed coin indicates the not valuable category. m is the number of dimensions that are modulated ($m = 3$) and n is the number of possible modulator dimensions ($n = 3$). For r and w the first number in the indices indicates the dimension, the second number in the indices indicates the feature. For the modulators v, the red numbers indicate which dimensions is modulated, the second number indicates which dimension is the modulator and the third number indicates the specific feature within that modulator.*

5.2.2 Rule Network

Like previous models, CAL assumes that stimulus dimensions are processed independently. In the humanoids' case each of the three varying dimensions (color, number and orientation) are treated as a separate dimension and their binary features are treated as nodes of each dimension. The set of dimensions is referred to as M, with m denoting each individual dimension (here: 1, 2 or 3). The set of features corresponding to a dimension is referred to as I with i denoting each individual feature (here: 1 or 2). In the current example, the humanoid with a blue hat, two stars on the belly and downwards pointing shoes is coded as follows: the blue hat is coded 11 (dimension 1, feature 1), the two stars are coded as 22 (dimension 2, feature 2) and the upwards pointing shoes are coded as 31 (dimension 3, feature 1). For each feature of the incoming stimulus (i.e. 11) associative weights w are computed for both the present and the absent category, in this case for both the valuable (i.e. $\mathrm{w11}$) and the non-valuable category (i.e. $\mathrm{w11}$). The associative weights to the valuable category are a product of how much attention was paid to the respective dimension, the information learned in the past about the feature via excitatory generalization (described above) and prior belief about the membership of that feature to the valuable category. Conversely, the associative weights to the not-valuable category are a product between how much attention was paid to the specific dimension, the information learned via inverse generalization about the feature and the prior belief on whether this feature belongs to the valuable category. At this stage, the model uses its first **free parameter γ**, which controls the extent of excitatory and inverse generalization. γ reflects how much information participants are actually generalizing and how refined this generalization is. In other words, if the information gradient from blue hat to green hat is sharp (which would indicate that blue hat belongs to the valuable category and the green hat belongs to the not-valuable category) or weak (it can barely be discriminated to which category blue hats or green hats belong to).

The associative weights are then used to compute dimension-wise evidence ratios:

$$r_{mI\bullet} = \ln\left(\frac{w_{mI\bullet} + .1}{w_{mI\otimes} + .1}\right), \quad (1)$$

where m and i take the above-mentioned values. The absolute value of r_{mI} indicates the strength of the evidence for a certain rule, and its sign reflects whether this evidence was for or against the rule (+ or - , respectively). As a last step, this product is multiplied by the appropriate contextual modulator. For example, it could be that, although in the past the blue (11) hat was predictive of the valuable category, together with two stars on the belly it led to error (the stimulus belonged to the not-valuable category). In this case, the model recognizes the two stars (22) as a contextual modulator $v_1$of the color dimension, dimension 1 (full notation being v_{122} as in **Figure 15**). This modulation is governed by the second **free parameter ω** which indicates the extent to which the participants are actually integrating the information from the context (i.e. whether they are able to recognize and incorporate the context in the prediction). The association between a rule prediction and a response is controlled by a gating mechanism z_{mIk} , where k takes the value of the possible responses (here ● or ⊗). First, each of these ratios is weighted by how much attention was paid to their respective dimension (e.g. r_{11} by α_1, r_{22} by α_2). Then, this product is adjusted depending on whether modulation occurs or not. If no modulation occurs, the weighted ratio is gated onto its corresponding response (e.g. if r_{mI}. indicates ●, z_{mIk} will also indicate ●). If modulation does occur, z_{mIk} weights r_{mI} by the corresponding modulator and gates it to the "opposite" response (i.e. ⊗).

5.2.3 Configural Memory

The architecture of the configural memory network is akin to the one in previous similarity-based models (e.g. GCM). The network responds to an incoming stimulus by computing psychological distances d_y between the current stimulus and previously stored stimuli (**Figure 15** first cube from below). These distances are used to calculate the strongest neighbor of the current stimulus. The strongest neighbor concept is similar to selecting a nearest

neighbor which is weighted by how strong this neighbor was encoded. This encoding strength is governed by the third **free parameter λ** which controls for the degree to which a stimulus is encoded and the ability to retrieve it (CAL assumes that encoding strength and retrieval are directly correlated). In the case illustrated in **Figure 15**, the strongest neighbor is the humanoid with a blue hat, one star on the belly and shoes pointing downwards (second cube from below). As in the rule network, associative weights between the selected stimulus and the valuable category ($a_{Y\bigcirc}$) and associative weights between the selected stimulus and the not-valuable category ($a_{Y\otimes}$) are calculated.

It has to be highlighted that CAL recruits the configural memory when the rule network fails to predict correct classification or when exceptions from the rule appear. It follows that in addition to prior knowledge about the stimulus, in configural memory the association a_Y of a stimulus to any category also highly depends on the current state of the rule network. Therefore, information on whether the current rule is successful or not and its evidence (the r ratio of the most attended to dimension) is also included in the calculation of $a_{Y\bigcirc}$ and $a_{Y\otimes}$. That is, if the rule network strongly predicts that "blue hats are valuable" and this prediction has been repeatedly correct, the configural memory network will learn very little about the current stimulus and the contribution to the overall predictions will be minimal. However, if the rule turns out to be incorrect, the memory network will learn a lot about the stimulus and will greatly contribute to the overall prediction. As in the rule network, an evidence ratio is computed but this time for the selected stimulus as a whole instead of its individual features:

$$H_{\bigcirc} = \ln\left(\frac{a_{Y\bigcirc}}{a_{Y\otimes}}\right) \tag{2}$$

5.2.4 Overall prediction

As mentioned above, CAL assumes that rule-learning and memorization are not separated, but interact and inform each other. This was reflected in the configural memory in the calculation of the associative weights a_Y. In the rule network, the interaction becomes apparent in the network's final prediction. The final prediction of the rule network $R'$is divided by the strongest prediction of the configural memory, such that:

$$R'_{\bullet} = \frac{1}{1 + \max(|a_Y|)} \sum_{m} z_{m\bullet} \quad (3)$$

The role of this division is to capture that when rule-learning is strong, a strong prediction from the memory network is indicative of an exception from the rule (note: exceptions are exceptions from both modulated and unmodulated rules).

Lastly, the predicted trial-wise probability that a stimulus S belongs to the valuable category is a logistic function containing the prediction from both the rule network R' and the configural memory network H as follows:

$$p(\bullet|S) = \frac{1}{1 + \exp(-2.5\,|R'_{\bullet} + H_{\bullet}|)} \quad (4)$$

It should be mentioned that for a two-stimulus display, these probabilities and their underlying calculations are computed for each stimulus on the screen, and then the stimulus with the highest probability is ultimately selected.

5.3 Challenges for the current paradigm

Two aspects were particularly challenging when adapting the model to the current paradigm. First, CAL was previously designed to accommodate one-stimulus display paradigms. The two-stimulus display raised questions on how the second stimulus would be best addressed, namely whether the updating process should take into account the stimulus position on the screen (left or right) or rather its category membership, since each pair display contained both a valuable and a not-valuable stimulus. Drawing inspiration from the reinforcement learning literature, the best option was found to be implementing a "fictitious RL" mechanism (Gläscher et al., 2009) which proposes separate learning rates for the chosen and unchosen option. Similarly, it was conceptualized that in the current paradigm participants would learn at different rates about the valuable and not-valuable stimuli. Instead of learning rates, in the current model this aspect was implemented at the level of generalization and contrasting, by "splitting" the corresponding γ parameter in two, a $\gamma_{valuable}$ and $\gamma_{not-valuable}$. As in Gläscher et al. (2009), differentiating learning based on

category membership captures participants' knowledge that in any given trial the category membership of the two stimuli was anticorrelated.

The second challenge was addressing the aspect that subjects were aware of the probabilistic nature of the feedback. As described in **Chapter 3.3.2**, participants were thoroughly instructed about the concept and performed training tasks to familiarize themselves with it. The option of introducing an additional "ignore feedback" parameter was considered, but this raised additional difficulties in setting suitable constraints for when this parameter should cancel the feedback-based updating (i.e. how many errors can be ignored so a rule is still considered correct). The alternative option was to suppose that under probabilistic feedback, rule-learning is more supervised. To capture this aspect, a "congruence check" between rule errors and the memory prediction was introduced. In other words, when the feedback suggests that the rule is erroneous, a check is run on whether this a true rule error that matches the memory prediction, and the respective item is indeed an exception stimulus (the memory network in CAL is assumed to be always "right"). Since CAL recruits the memory network when rules are erroneous, this mechanism prevents wrong updating by double-checking if the respective stimulus was indeed a stimulus strongly stored in memory (i.e. an exception from the rule). It has to be highlighted that this "congruence check" acts mostly at the rule-learning stage, and thus is mostly beneficial for the rule-based task. It was considered that, since memorization always tends to the true category in CAL (which would be the case even in probabilistic feedback), no equivalent was needed for the stimulus-based task.

5.4 Hypotheses

As mentioned previously, one of the main goals of applying CAL to the current data was to explore model-based differences between the two tasks which could potentially complement the understanding of the model-free behavioral differences. The primary target for assessing these differences were the four free parameters. For a better overview, these parameters and their associated cognitive functions are summarized in **Table 3**.

Table 3

CAL's free parameters and their associated cognitive functions.

Parameter	Main Role	Cognitive Functions	Interpretation
$\gamma_{valuable}$ (gamma valuable)	Similarity / Dissimilarity	generalization of valuable category (and modulator)	larger values indicate wider generalization gradients
$\gamma_{not-valuable}$ (gamma not - valuable	Similarity / Dissimilarity	generalization to not-valuable category (and modulator)	larger values indicate wider generalization gradients
λ (lambda)	Configural Memory (extent of encoding)	- encoding strength - moderates encoding of exceptions	larger values indicate stronger memory
ω (omega)	Contextual Modulation (sensitivity to context)	- extent of information-integration from context	larger values, stronger contextual modulation

Adapted from Schlegelmilch et al. (2021)

It was expected that the four parameters will differ between the two tasks as follows:

H1: $\gamma_{valuable}$ values will be lower in the rule-based task than in the stimulus-based task. This difference is expected due to the rule-like instructions in the rule-based task which should lead to narrower generalization gradients (stronger rule-learning).

H2: The λ values will be higher in the stimulus-based task than in the rule-based task, reflecting that participants used a memorization-based strategy in the former.

H3: Since the rule-based task can be solved by XOR rule, which is essentially a uni-dimensional rule which requires contextual modulation, the ω values will be higher in this task than in the stimulus-based task which should entail little to no contextual modulation.

H4: There will be a significant difference between $\gamma_{valuable}$ and $\gamma_{not-valuable}$, reflecting the assumed learning difference between valuable and not-valuable stimuli.

5.5 Materials and Methods

For both the rule-based and stimulus-based task, the model was fitted to individual data (N = 30). The best-fitting parameters were estimated using a differential evolution (DE) optimization method as implemented in the R package DEoptim (Ardia et al., 2015). Being an evolutionary algorithm, DE "optimizes a problem by iteratively improving a candidate solution based on an evolutionary process" (Georgioudakis & Plevris, 2020). 500 iterations were used. In the original manuscript, this optimization was proven superior to parameter grid searches or gradient-based methods (Schlegelmilch et al., 2021).

CAL performance was assessed against a random agent model. Trial-wise log-likelihoods were computed separately for each fitted participant in the rule-based task and stimulus-based task. Trials in which participants failed to respond were not included. For the random agent model, the log-likelihoods were computed by assuming constant random choice, in other words, a predicted probability of 0.5 % in each trial. In order to obtain a more accurate model comparison, in the log-likelihood calculation for each assumed random agent the number of trials was matched to the raw data. For example, if one participant provided a response in 156 trials, its corresponding random agent was calculated assuming 156 trials. These calculations were done separately for the rule-based random agent and stimulus-based random agent. The CAL mean log-likelihood of each task was compared with the mean log-likelihood of the corresponding random agent model.

5.6 Results

CAL's model fits for the two tasks outperformed by far those of the random agent model (*CAL log-likelihood*$_{rule\text{-}based}$ = -37.45; *random agent log-likelihood*$_{rule\text{-}based}$ = -110.30; *CAL log-likelihood*$_{stimulus\text{-}based}$ = -51.19; *random agent log-likelihood*$_{stimulus\text{-}based}$ = -110.14). **Figure 16** contains examples of fitted participants. The CAL curves indicate the match between the CAL prediction and the participant's response, with 1 indicating a perfect match. It can be observed that for the well-fitted participants (**Figure 16** top) CAL managed to capture not only most of the correct responses but also most of the errors that

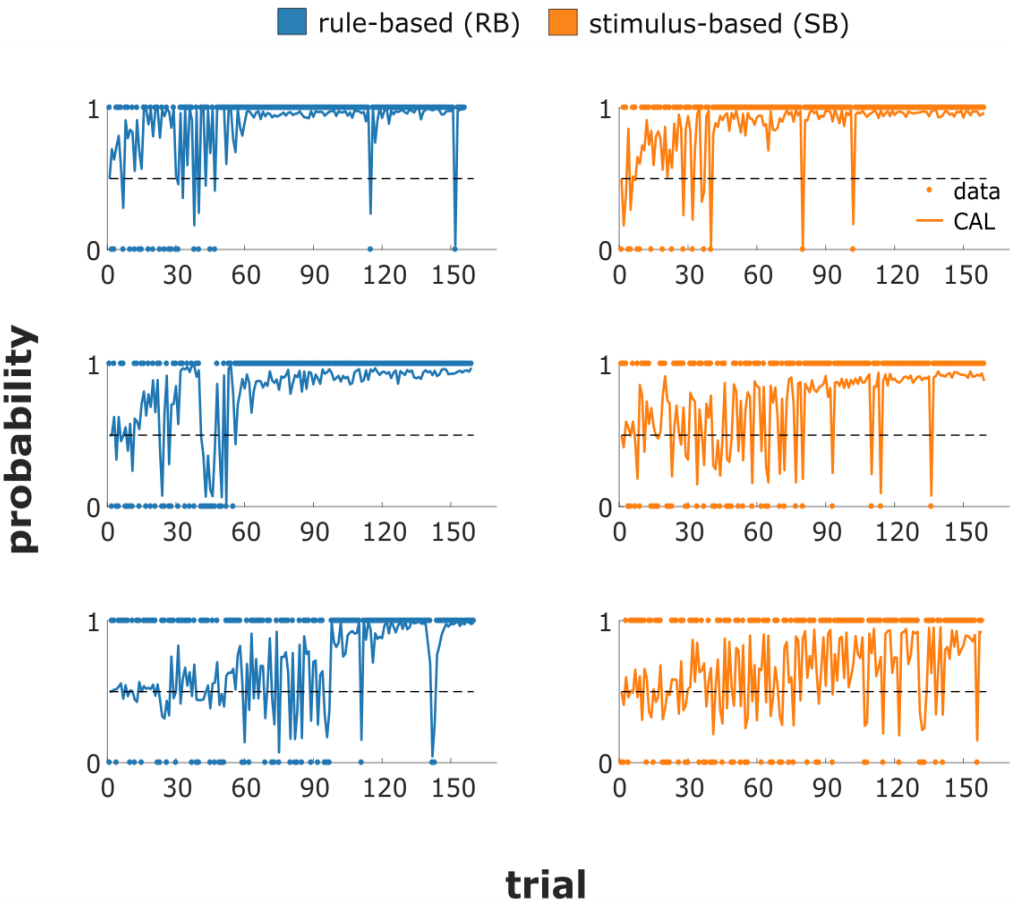


Figure 16. *Example CAL predictions. Dots represent participants' choice data where 0 indicates an incorrect response and 1 indicates a correct response. Solid lines indicate the fit between CAL's prediction where 0 indicates no fit between the model prediction and participant's response (i.e. the model predicted a correct response although the participant answered incorrectly) while 1 indicates a perfect fit (i.e. the model predicted a correct response and the participant responded correctly). Dashed lines indicate chance level.* ***Left****: accuracy and fit of three subjects in the rule-based task.* ***Right****: accuracy and fit of three subjects in the stimulus-based task. Both sides indicate a participant whose data was fit well by the model (top), a participant whose data was moderately fit (middle) and a participant whose data was poorly fit (bottom).*

participants made. The model only failed to predict three errors up to learning and two errors towards the end of the task, which were likely due to fatigue. For the participants with moderate fits (**Figure 16** middle), the model was not able to capture the errors for a certain period during learning. For example, for the rule-based participants, some of the errors made between trials 50 and 70 were not captured by the model. Nevertheless, the predictions quickly improved and the model managed to "learn" at approximately the same time as the participant.

As far as the group performance is concerned, on average CAL fitted the rule-based data slightly better than the stimulus-based data, having an overall accuracy of 84 % (*SE* = 14.95 %) for the participants in the rule-based task as opposed to 79 % (*SE* = 15.46 %) in the stimulus-based task. The inferior fit of the stimulus-based task becomes apparent when looking at the block-wise raw and predicted accuracies in **Figure 17**. It can be observed, that the model overestimated performance in the stimulus-based task. In the data, the stimulus-based performance after learning (block 5 to block 10) stayed below the one in the rule-based task and never reached 1. By contrast, the predicted stimulus-based performance matched the post-learning performance in the rule-based

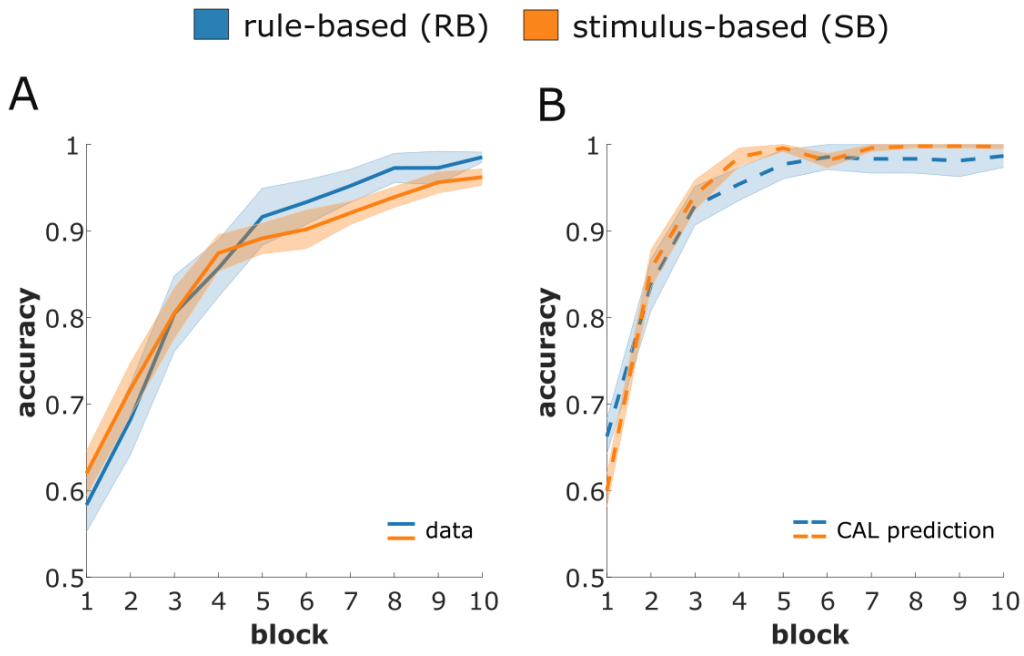


Figure 17. *Comparison between group data and CAL predictions.* ***A.*** *Block-wise mean accuracy of each tasks' raw data (N = 30).* ***B.*** *CAL's predicted block-wise accuracy. A block consists of 16 trials. Shaded areas indicate standard errors.*

task (and even slightly exceeded it) and stabilized around 1 in the last blocks. Nevertheless, CAL successfully predicted that the rule-based task and the stimulus-based task were learned at approximately the same time.

To get a better understanding of CAL's predictions, it was assessed how the models' accuracy correlates with each participant's behavior. For both tasks, there was a strong positive correlation between the model and participants' mean overall accuracy, both in the rule-based task and the stimulus-based task (*Spearman rho's* $_{rule\text{-}based}$ = 0.93, p < .001; Spearman rho's $_{stimulus\text{-}based}$ = 0.85, p < .001). The opposite relationship was found between the model accuracy and participants' individual learning points (as described in **Chapter 3.4** and displayed in **Figure 8**). There was a strong negative relationship between model accuracy and learning points. However, this correlation was stronger for the rule-based task than for the stimulus-based task (*Spearman rho's* $_{rule\text{-}based}$ = -0.81, p < .001; *Spearman rho's* $_{stimulus\text{-}based}$ = -0.64, p < .001).

Next, the first three hypotheses were investigated, which proposed task-related differences between the fitted parameters. **Figure 18** contains the distributions of the best-fitting parameters in each task condition. Differences between the two tasks could be observed in the $\gamma_{valuable}$ and ω parameters. Statistical analyses confirmed these observations. The differences in $\gamma_{valuable}$ were assessed with a paired samples *t*-test which indicated a significantly smaller mean $\gamma_{valuable}$ in the rule-based task ($M_{rule\text{-}based}$ = -1.05, *SE* = 0.2) than in the stimulus-based task ($M_{stimulus\text{-}based}$ = -0.03, *SE* =0.31, *t(29)* = -2.4, p = 0.02). Although the $\gamma_{not-valuable}$ values suggested a similar pattern, their difference was minimal and not statistically significant. A Wilcoxon signed-rank test revealed that the ω values were significantly lower for the rule-based participants ($Mdn_{rule\text{-}based}$ = -1.36) as opposed to the stimulus-based participants ($Mdn_{stimulus\text{-}based}$ = 0.32, Z = -2.09, p = .03). With respect to the λ parameter, although the values in the stimulus-based task were higher than in the rule-based task, their difference did not reach significance ($Mdn_{rule\text{-}based}$ = 0.56, $Mdn_{stimulus\text{-}based}$ = 1.63).

The last hypothesis, which posed a distinction between the two generalization parameters γ, was evaluated using Wilcoxon signed-ranks tests. The results revealed that in both tasks the median $\gamma_{valuable}$ values were smaller

than $\gamma_{not-valuable}$ values, but this distinction only reached significance in the rule-based task (*Mdn*$\gamma_{valuable}$ = -0.97, *Mdn* $\gamma_{not-valuable}$ = 1.61, *Z* = -4.47, *p* < .001).

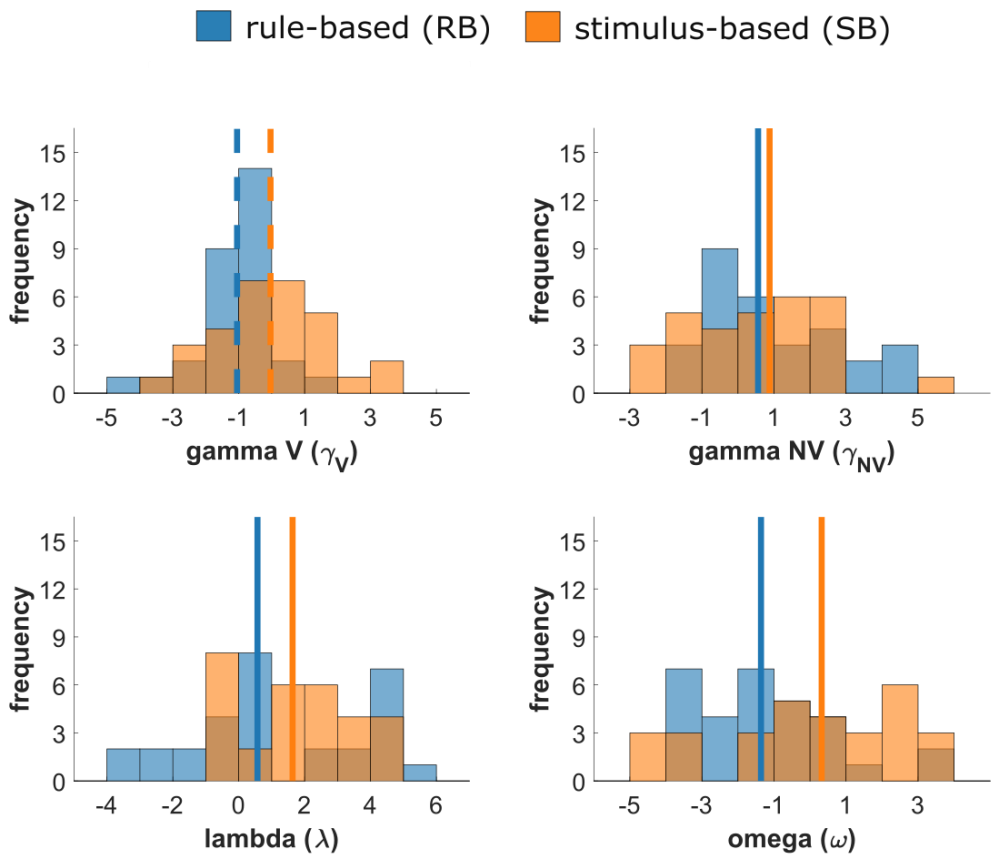


Figure 18. *Histograms of estimated parameters. Dashed lines indicate group means while continuous lines indicate group medians. The color of each line indicates whether it is representative of the rule-based or stimulus-based task.*

Next, it was explored whether there was a relationship between the fitted parameters and participants' performance. **Figure 19** displays the individual estimated parameters plotted against the mean accuracy of each participant. The λ plot suggests a very strong positive correlation between the lambda values and mean accuracy in both tasks. A Spearman's rho correlation confirmed this observation ($rho_{rule\text{-}based}$ = 0.91, *p* < .001; $rho_{stimulus\text{-}based}$ = 0.83, *p* < .001). As far as the other parameters are concerned, a correlation was found between ω values and accuracy in the rule-based task only ($rho_{rule\text{-}based}$ = 0.62, *p* < .001). With respect to the generalization parameters, as indicated in the plot, $\gamma_{valuable}$ did not correlate with accuracy and $\gamma_{not-valuable}$ showed only a weak

correlation with the performance in the stimulus-based ($rho_{stimulus\text{-}based}$ = 0.38, p = .03).

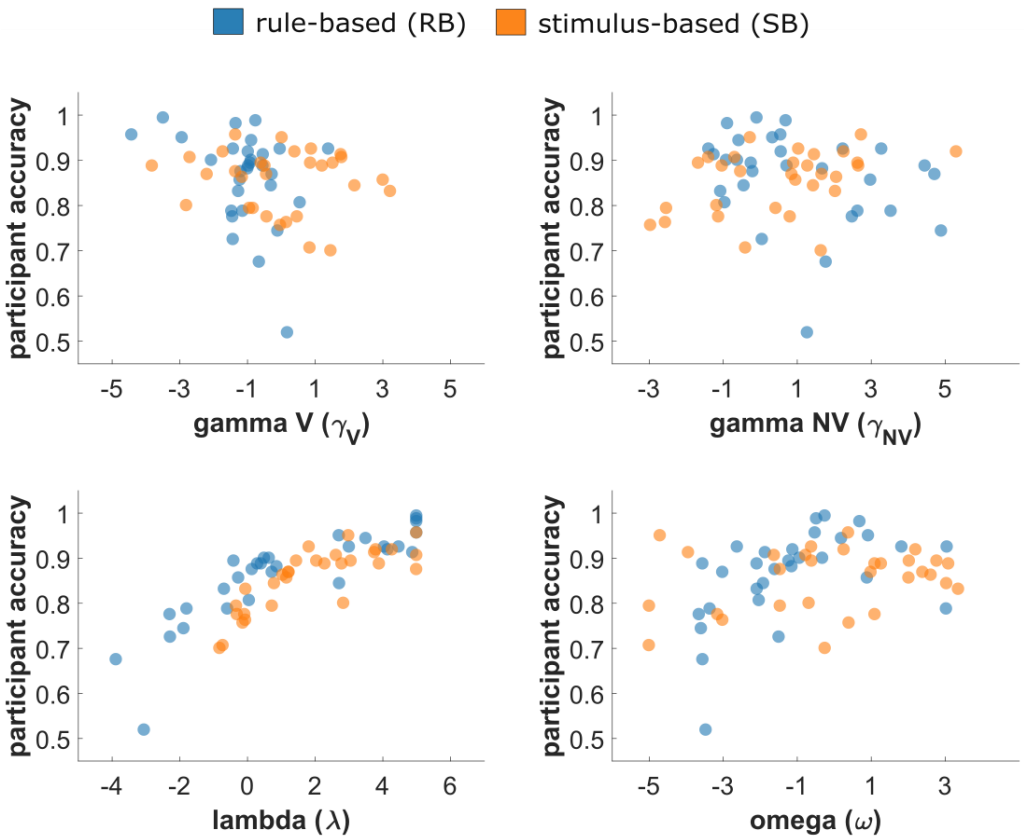


Figure 19 *Correlations between CAL parameters and participants' accuracy. X-axes indicate individual participants parameter value. Y-axes indicate participants mean accuracy in proportions (1 = perfect accuracy).*

5.7 Discussion

This chapter described the first attempt to model the two-stimulus display adaptation of type II and type VI problems using the newly developed CAL model. The model was the strongest candidate for the current work not only owing to its ability to capture previously unaddressed shortcomings, but also due to its networks mirroring the two investigated learning types: rule-based and stimulus-based. Since this was a pioneer work, it was first assessed whether the model was able to fit each task well, and whether it captured the performance differences between the two. CAL's accuracy exceeded 75% on both tasks, and more importantly, was able to capture the lack of type II

advantage in the group performance. The model did however, predict slightly steeper learning in both tasks compared to the raw data and overestimated the group performance in the stimulus-based task.

While a larger sample size would have quite likely led to a higher model fit, some qualitative explanations can also be proposed. As far as the rule-based task is concerned, the fast learning could be explained by the effect of rule-like instructions proposed by Schlegelmilch et al. (2021). Specifically, in CAL, rule-like instructions affect rule-learning by leading to sharper generalization gradients, which in turn quickly activate contextual modulation. Given the rule-like instruction of the current task, and the found distribution of the ω values, this assumption could already explain the predicted fast learning in the task. The slight mismatch between the predicted mean accuracy and raw data could be due to the implemented "congruency check" (described in **Section 5.3** of this chapter) being highly efficient. It is plausible that participants' "congruency check" is less optimal and that perhaps they did not double-check in their memory after every single rule error.

With respect to the stimulus-based task, the overestimation of accuracy could be attributed to either the probabilistic feedback or the instruction type. CAL starts by trying to form simple rules, and when these rules fail it recruits the configural memory. Without taking the instructions into account, the structure of the stimulus-based task (in which simple rules fail) together with the probabilistic feedback quite likely resulted in a considerable amount of negative feedback early in the task. Therefore, the participants could have switched to the configural memory early in the task, and thus quickly became very efficient. A similar argument can be made about the instruction type, which is that participants started memorizing early on, and consequently the memory network made strong predictions early on. Nonetheless, in order to disentangle these mechanisms more research is needed. In particular, future studies could attempt to simulate the type VI problems or the stimulus-based task under different experimental conditions (e.g. with and without instruction, with and without probabilistic feedback).

This chapter also investigated whether CAL's free parameters (summarized in **Table 3**) would give insights into the differences between the

two tasks. Although a core assumption of CAL is that both networks are highly interconnected and inform each other, the generalization parameters γ and the contextual modulation parameter ω play a stronger role in the rule network. Thus, it was expected that these two parameters would reveal important aspects about the rule-based task. Conversely, the "main" parameter in the configural network, which governed the encoding strength, was expected to have higher values in the stimulus-based task.

Concerning the rule network parameters, the γ_{valuable} findings aligned with the expected differences. Aside from a reassurance that two distinct strategies were used in the two tasks, the lower γ_{valuable} values in the rule-based task are an indirect argument that most participants showed strong rule-learning (which requires sharp generalization gradients). However, the findings with respect to ω values hint that this argument should be regarded with caution.

Contrary to the initial hypothesis, the ω values were lower in the rule-based task than in the stimulus-based task. This was surprising since the XOR nature of the rule-based task was expected to elicit high contextual modulation. Although puzzling, one potential explanation could be found when looking at the training tasks. In the deterministic rule-based training task (described in detail in **Chapter 3.3.2**), participants were instructed that one dimension was irrelevant (they were also told which one). Although the main task instructions did not explicitly say that only one dimension would be irrelevant (participants were told that "some dimensions could be irrelevant") participants might have extrapolated from the training task, and could have expected an XOR rule. Therefore, participants could have needed less information from the context. Nonetheless, the high ω values in the stimulus-based task remain difficult to interpret. It could be the case that this is a previously unstudied effect of memorization instruction in type VI problems. On the other hand, one could argue potential transfer effects from the rule-based task in some participants. Nonetheless, owing to the minimum two-day time gap between the two tasks, this explanation is rather unlikely. It remains a possibility that despite the instructions discouraging rule-like learning, some participants could have attempted to find a rule solution. When asking participants to report how they solved the six problems, Shepard et al. (1961) found that some participants

reported rule-like solutions to type VI problems. Although the high stimulus-based λ values from the configural memory network (discussed below) refute this idea as a sole explanation, it cannot be excluded that a certain degree of rule-learning could have occurred in some participants at certain stages of the task. For example, some participants could have entertained a rule-like solution at the beginning of the task, before they realized that this strategy was unsuccessful. Another possibility could be the usage of rules as a "cognitive shortcut" at the end of the task after the stimuli have already been memorized. This shortcut could have been taken either due to boredom in the early learners or because certain participants, despite using the memorization strategy as instructed, generally found rule-application easier, and therefore applied it after learning had been completed. Nonetheless, no definite conclusion can be drawn without an empirical measure of contextual modulation. It is encouraged that future studies would attempt to develop a direct measure of this mechanism.

Regarding the configural memory network, the λ values were higher in the stimulus-based task than in the rule-based task, but this comparison did not reach significance. The high λ values in the stimulus-based task, combined with the high γ_{valuable} values, speak strongly in favor of a general memorization-based strategy. To reiterate, high γ_{valuable} values in the stimulus-based task reflect wide generalization gradients which can be interpreted in this case as weak or absent rule-learning, and thereby memorization. Moreover, the strong correlation between the λ values and the high accuracy in the stimulus-based task have a straightforward explanation: the stronger the encoding and retrieval, the better the memorization, hence, the higher the performance.

The finding that rule-based performance also positively correlates with λ values, in addition to certain participants having high λ values, are less straightforward to interpret. It has to be noted that CAL does not assume that strong encoding strength is anticorrelated with rule-learning. Instead, strong encoding strength is also regarded as a strong ability to retrieve stimuli that are exceptions from the rules – which could still indicate a potentially successful rule-based strategy. Furthermore, considering the probabilistic feedback, it is reasonable that subjects could have stored items in memory which were only temporarily regarded as a rule exception. Nonetheless, the high λ values could

also be explained by certain participants temporarily ignoring the rule-like instructions. This process could be either voluntarily due to participants trusting their memorization abilities more than their pattern-finding abilities or involuntary due to an uncertainty state induced by the probabilistic environment or simply because despite instructions, memorization could just not be "turned off".

The current application of CAL adapted to the two-stimulus display by employing two generalization gradients, $\gamma_{valuable}$ and $\gamma_{not-valuable}$ as opposed to only one γ, as it was initially conceptualized. The underlying motivation was that the participants would learn differently about the valuable and not-valuable members, especially given that only the valuable stimuli were rewarded. The findings were in line with the idea of category-dependent difference in learning, with $\gamma_{valuable}$ values being smaller than the $\gamma_{not-valuable}$ in both task conditions. This result suggests the fact that only the valuable stimuli were rewarded was reflected in sharper generalization gradients. In other words, participants abstracted information more efficiently in the valuable stimuli.

It has to be mentioned that this chapter presented just a small step in modeling the current task with CAL. Future work will attempt to directly compare performance of CAL on the current paradigm with previously established models such as SUSTAIN. More attempts will be made to better capture the effect of probabilistic feedback in both computational and cognitive sound manners. For example, inspiration can be drawn from a recent model-based EEG study of probabilistic categorization tasks (Sewell et al., 2018), which modeled the degree of feedback discounting in each participant. Furthermore, potential ideas for improvement can be found in CAL's predictions about attentional allocation. It is quite likely that the assessment of the two sets of attentional parameters also in combination with high-quality neural data would unravel new aspects about the two tasks that could be implemented in CAL to obtain a better fit.

All in all, using the new CAL model brought this work a step closer to understanding the cognitive computations behind the performance in the two tasks. CAL revealed that for both rule-based and stimulus-based learning, high encoding and retrieval strength are advantageous. Furthermore, in both tasks

participants assign more weight to the valuable stimuli than to the not-valuable stimuli, mechanism implemented by a sharpening of the generalization gradients. Overall, rule-based strategies are characterized by narrower generalization gradients than stimulus-based strategies.

While this chapter focused on the explanatory power of CAL's free parameters, the next chapter investigated what neural questions CAL's trial-wise estimates can answer about the two problems.

6. Dissociable Neural Signatures

6.1 Introduction and Hypotheses

The existence of multiple category learning systems is still controversial. Ever since its first formulation, the propositions of the COmpetition between Verbal and Implicit model (COVIS) have been fundamental in testing whether verbal (declarative) and implicit (procedural) category learning rely on multiple behavioral and neural systems (Ashby & Valentin, 2017). The most common approach in these assessments was the usage of rule-based tasks and information-integration tasks as means to compare declarative and procedural learning. Although considerable behavioral work has been done recently on investigating the multiple systems (or lack thereof) (Ashby & Valentin, 2016; Casale et al., 2012; Donkin et al., 2015; Edmunds et al., 2018; Smith et al., 2014; Smith & Church, 2018; Wills et al., 2019), fMRI work has been comparatively scarce (some of the most recent ones being Milton & Pothos, 2011; Waldschmidt & Ashby, 2011). Evidence from this fMRI work leaned towards a consensus that the two types of learning involve two separate neural systems such that declarative category learning is characterized by prefrontal and medial temporal role activation, while procedural category learning is strongly striatal-based. Nevertheless, recent behavioral studies have brought strong criticism to the paradigms used in these fMRI studies, such as erroneous identification of the underlying strategies and extrapolation from group data that averages over strategy subgroups (Edmunds et al., 2018). In light of this evidence, the findings suggesting two dissociable neural systems are called into question.

Given the concerns raised about the employed paradigms, in addition to the lack of recent fMRI studies, it is argued that the neuroscientific community could benefit from new fMRI perspectives. Accordingly, this work took a different approach and investigated the neural correlates of declarative and procedural categorization by using a two-stimulus adaptation of Shepard's type II and type VI problems. Due to their underlying optimal strategies, i.e. rule-like solution as opposed to non-verbalizable solution, the two problem types were treated as proxy for the two learning types. It is worth mentioning that although participants might be using an exemplar-based approach in the current adaptation of type

VI problem, this approach has been previously regarded as an equivalent for the implicit system in COVIS (Erickson & Kruschke, 1998; Pickering, 1997 as cited in Ashby et al., 2003).

To date, there have only been two other fMRI studies addressing Shepard's type II and type VI problems. Mack et al. (2016) used Shepard's type I and II problems to study how new concepts are represented in the hippocampus, and whether this representation changes dynamically as a function of task goal. The authors found that switching from type I problem to type II problem and vice versa correlated strongly with hippocampal involvement, representing goal-related concept update. Moreover, the updating process itself was associated with a coupling between ventromedial prefrontal cortex (vmPFC) and hippocampus. In a follow-up paper, Mack et al. (2020) employed type I, II and VI problems to test whether the vmPFC adapts to the changing goals by "compressing" information irrelevant to the task at hand. This dynamic reduction hypothesis was confirmed by a strong correlation between vmPFC and problem complexity. The vmPFC had the highest compression score in the type I problem which required the most information reduction (only one dimension has to be attended to solve the task), and the lowest compression score in the type VI tasks which required the least information reduction (all features have to be attended to solve the task). Two aspects are worthy of note here. First, the focus of the studies was on the effects of task-goal changes and not on the problems per se. Second, in both studies, no clear instructions were given on how to perform the task (in fact no study up to date has instructed participants on how to perform type VI), and therefore it is likely that the observed activity corresponded not only to procedural or declarative learning but also to additional processes (i.e. search for appropriate strategy). By contrast, the current work focused on the differences between the two problem types and their unique neural signatures. It was intended that through strategy instructions encouraging the corresponding optimal solution, effects of goal or strategy changes would be reduced, and thus a cleaner picture of the differences between the two would be obtained.

An important common aspect of the two studies mentioned above was the use of model-based fMRI. This technique has been popular in decision

making literature (Bayer et al., 2020; Doll et al., 2012; Gläscher et al., 2010; Gläscher & O'Doherty, 2010), and has been gaining attention in the category learning literature (Davis et al., 2012a, 2012b). In short, while conventional fMRI analyses reveal the involvement of a certain region in a cognitive task or condition, model-based fMRI allows testing of more specific hypotheses about the relationship between the cognitive processes and neural activity (Wilson & Niv, 2015). This is done by fitting a suitable cognitive model to the behavioral data and then correlating the latent variables of interest with the neural activity. The model of choice in Mack et al. (2016) and Mack et al. (2020) was SUSTAIN which was particularly beneficial for investigating the goal-related attentional changes of interest. By contrast, the current work focused on the underlying learning strategies per se. Thus, the neural activity was correlated with the trial-wise predictions of rule-learning and configural memory derived from the newly developed CAL model described in **Chapter 5.2**. It was aimed that by taking this detailed approach, the computations carried out by each region would become clearer, and would help elucidate whether they are reminiscent of one or two category learning systems.

Since the neural correlates were investigated with both conventional analyses (referred henceforth as model-free) and model-based approaches, separate hypotheses were formulated for each approach.

6.1.1 Model-free hypotheses

In light of the past research two model-free hypotheses were formulated:

H1: The rule-based task will elicit more activity in the prefrontal cortex and hippocampus than the stimulus-based task.

H2: The stimulus-based task will engage striatal regions more strongly than the rule-based task.

6.1.2 Model-based hypotheses

Given that this is the first model-based fMRI application of the CAL model, the current work was mainly exploratory. It was expected that the two CAL networks, the rule network and configural memory (described in detail in

Chapter 5.2) would differentially correlate with brain activity in the rule-based and stimulus-based task such that:

H1: Activity in the prefrontal cortex and hippocampus will correlate more strongly with the trial-wise estimates from the rule network than activity in striatal regions.

H2: Activity in the striatal regions will correlate more strongly with trial-wise estimates from the configural memory network than the activity in the prefrontal cortex and hippocampus.

6.2 Materials and Methods

6.2.1 Participants

Of the 30 participants that successfully completed behavioral tasks, three participants had to be excluded from the fMRI analyses due to technical problems during data acquisition. The remaining sample contained 8 males and 19 females (M_{age} = 25.5, SD_{age} = 2.77).

6.2.2 Data Acquisition

Event-related fMRI data was collected using a 3T Siemens Scanner (Siemens, Erlangen, Germany) with a 32-channel head coil. An MR-compatible mirror for eye-tracking recordings was attached to the coil so that participants could see the mirrored image of the MR compatible screen. The fMRI images were acquired using multiband gradient Echo-Planar Imaging (EPI) with a TR of 1636 ms, TE of 29 ms (FoV = 224 mm, flip angle = 70 degrees, multiband factor = 2). Each volume contained 54 slices, with a 2x2x2 voxel size and 2 mm slice thickness. A field map was acquired on each testing day after the localizer (TE1= 5.51 ms, TE2 = 7.91 ms). At the end of Day 2, a high resolution anatomical T1-weighted image (MPRAGE protocol) and a medial temporal lobe T2-weighted image (28 slices, TR = 9520 ms, TE = 80 ms) were acquired. The physiological responses (respiration and skin conductance) measured during both scanning sessions with a Biopac MP 100 (Biopac Systems, Inc) are not part of this manuscript.

6.2.3 Preprocessing

Data was pre-processed and analyzed using the SPM12 software (Welcome Department of Cognitive Neurology, London, UK) running under Matlab R2014b (Neurobehavioral Systems, Inc., Berkeley, CA, USA). The first five images from each scanning session were discarded in order to avoid spin saturation confounds. In the first pre-processing step, the day-specific field map was applied to the functional volumes to correct for possible magnetic field inhomogeneities. No slice timing correction was performed due to the small TR of 1636 ms. Next, the functional images were realigned and unwarped to remove susceptibility-by-movement artifacts. The resulting images were coregistered to their corresponding subject-specific anatomical images (T1-weighted). Anatomical images were segmented into gray matter, white matter and cerebral spinal fluid probability maps. Based on these maps, the "Diffeomorphic Anatomic Registration Through an Exponentiated Lie Algebra" (DARTEL) SPM toolbox was used to normalize the images into the Montreal Neurological Institute (MNI) standard stereotaxic space. In the last pre-processing step, all functional images were smoothed using a Gaussian kernel with a full width half maximum of 6 mm to increase the signal-to-noise ratio and fulfill the requirement of random Gaussian field theory.

6.3. Results

6.3.1 Model-free fMRI

6.3.1.1 Statistical analyses

The model-free hypotheses were investigated using a general linear model as implemented in SPM using a mass univariate approach. This approach entails a first-level analysis, in which the subjects are modeled individually, and a second level-analysis in which the group data is modeled.

Within the first-level analysis, subject-level models are defined which include the predictors of interest. The dependent variable in this analysis is the activity within a single voxel. The within-subject character of the study facilitated the inclusion of both tasks within the same model, such that the two different

tasks were regarded as two conditions of the same task. Within each task, a distinction was made between the choice period of the task and the feedback period of the task. Importantly, the full choice period was included in the model irrespective of when participants made a choice. The trials in which participants failed to respond were excluded from the analysis. Thus, the predictors of interest were: rule-based choice, rule-based feedback, stimulus-based choice, and stimulus-based feedback. In order to increase the specificity of the model, each choice regressor was parametrically modulated by the respective choice, and each feedback regressor was parametrically modulated by the accuracy. No nuisance regressors were included in the model. The regressors of interest and their corresponding parametric modulators were subsequently convolved with the canonical hemodynamic response function (HRF). Serial correlations were accounted for using the SPM12 'FAST' method. For all other estimation parameters, the default values from SPM12 were used.

The second-level analysis was based on subject-specific beta images from the rule-based choice and stimulus-based choice regressors. The first contrast tested for regions that are more active during the rule-based task as opposed to the stimulus-based task. Thus, the rule-based choice beta estimates were assigned a weight of 1, and the stimulus-based choice beta estimates were assigned a weight of -1. The second contrast tested the reverse differences, namely which regions are more active during the stimulus-based task than in the rule-based task, and therefore the rule-based choice betas were assigned a weight of -1 and the stimulus-based choice betas were assigned a weight of 1.

All contrasts of interest were assessed using one sample *t*-tests. Results were family-wise error corrected for multiple comparisons within specified anatomical masks. All anatomical masks were retrieved from Harvard-Oxford cortical and subcortical structural atlases (www.fmrib.ox.ac.uk/fsl) except for the vmPFC mask which was retrieved from a previous study (Clithero & Rangel, 2014). All results were considered significant at $p_{FWE} < .05$.

6.3.1.2 Results

The first model-free hypothesis was addressed by contrasting the activity during choice in the rule-based task to the activity during choice in the stimulus-based task within the two regions of interest. This contrast (**Figure 20A)** revealed strong bilateral hippocampal involvement (peak voxel $_{x, y, z}$ = 24, -10, -18, Z = 5.49, p_{FWE} < .0001) as well as anterior and central vmPFC (peak voxel $_{x, y, z}$ = -4, 52, -2, Z = 5.57, p_{FWE} < .0001) (**Table 4**). The corresponding mean first-level beta estimates are presented in **Figure 20B**. It can be seen that the

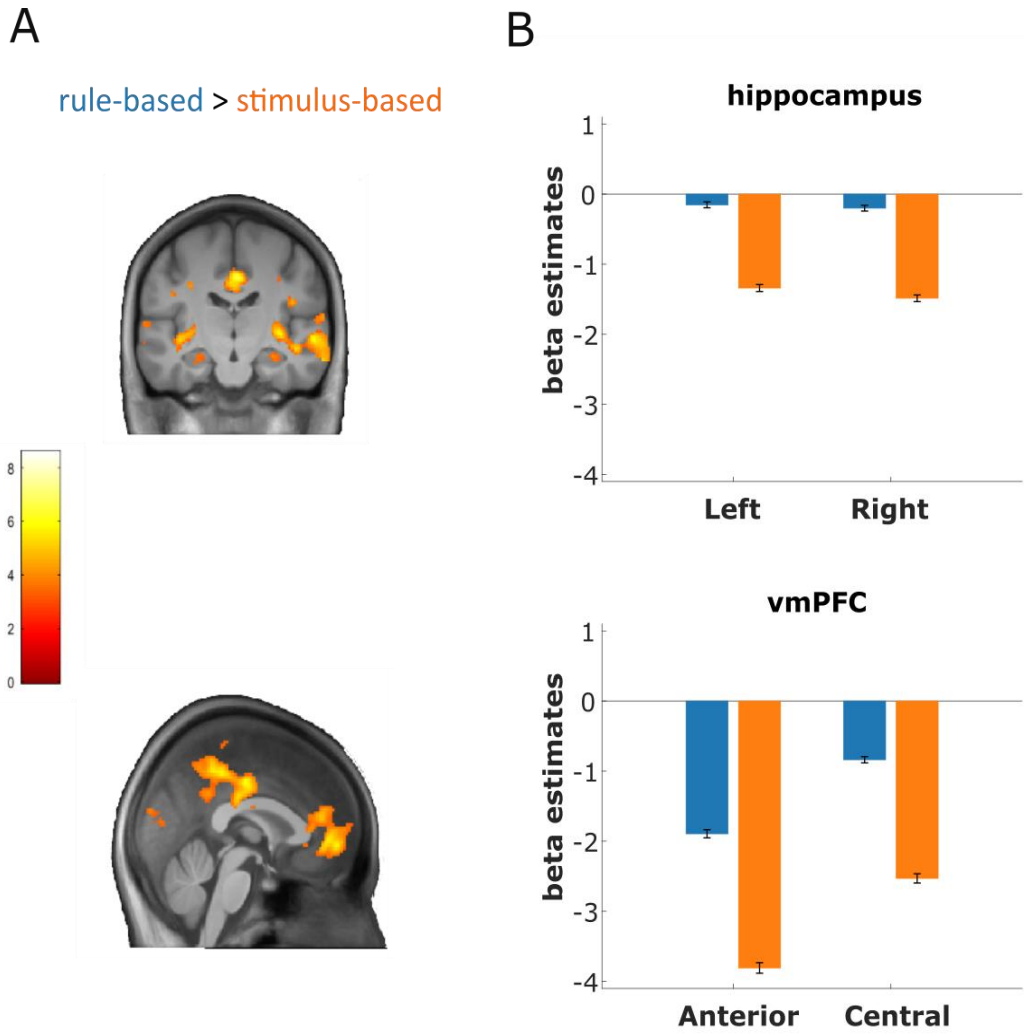


Figure 20. *Group model-free fMRI results (N = 27).* ***A.*** *Second-level contrast rule-based > stimulus-based. Brighter regions indicate higher t-values. Results thresholded at p < .001 for visualization purposes. Both statistical maps are overlaid onto the group anatomical image (mean T1). Color scale indicated t-values.* ***B.*** *First-level beta estimates (in arbitrary units) for the peak voxel in the respective region of interest. Blue bars indicate mean betas for the rule-based choice regressor. Orange bars indicate mean betas for the stimulus-based choice regressor extracted from peak voxels (coordinates listed in* ***Table 4****). Error bars indicate standard errors.*

two tasks were characterized in both regions by a decrease with respect to the model's implicit baseline, with the stimulus-based task decreasing more drastically than the rule-based task (all paired *t*-tests within each region significant at $p < .001$).

Table 4

Group fMRI results for the rule-based > stimulus-based t-contrast.

Region	Side	Peak voxel MNI coordinates x y z (mm)	Cluster size	Z score
Hippocampus	left	-32 -30 -12	51	5.06
	right	24 -10 -18	105	5.49
vmPFC	anterior	-4 52 -2	96	5.57
	central	-4 46 2	157	5.18

The second model-free hypothesis posing that the stimulus-based task will correlate more strongly with striatal activity than the rule-based task was tested using the reverse contrast within the ventral striatal regions. Compared to the rule-based task, the stimulus-based task elicited higher bilateral caudate (peak voxel $_{x, y, z}$ = 12 16 0, Z = 5.62, p < .0001) and putamen activity (peak voxel $_{x, y, z}$ = -16 10 -4, Z = 4.78, p < .0001) (**Figure 21A**). Exploratory analyses indicated also a bilateral insular contribution (peak voxel $_{x, y, z}$ = 34 22 2, Z = 5.48, p < .0001). The corresponding beta values (**Figure 21B**) suggest that both tasks were characterized by an increase with respect to the implicit baseline, with the

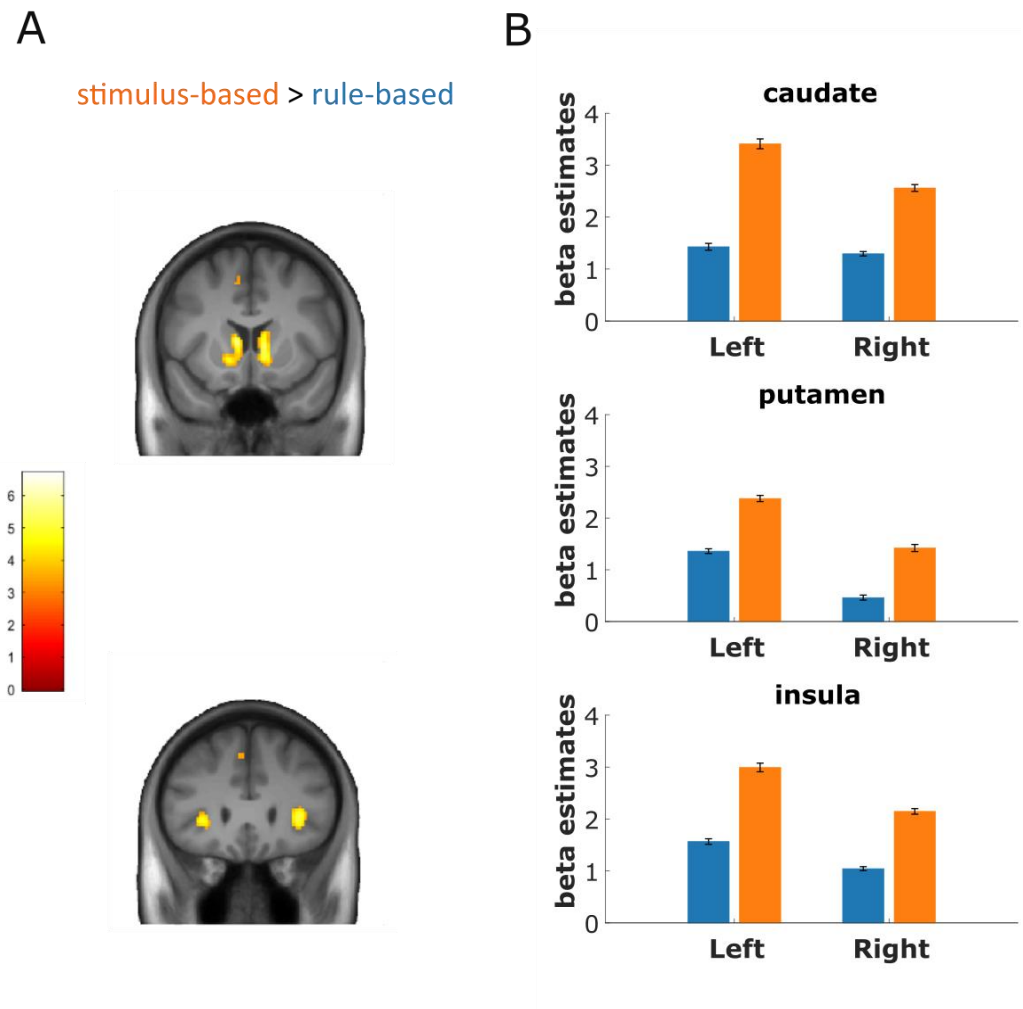


Figure 21. *Group model-free fMRI results (N = 27).* ***A.*** *Second-level contrast stimulus-based > rule-based. Brighter regions indicate higher t-values. Results thresholded at p < .001 for visualization purposes. Both statistical maps are overlaid onto the group anatomical image (mean T1).* ***B.*** *First-level beta estimates for the peak voxel in the respective region of interest (**Table 5**). Blue bars indicate mean betas for the rule-based choice regressor. Orange bars indicate mean betas for the stimulus-based choice regressor. Error bars indicate standard errors.*

stimulus-based task increasing more than the rule-based task in all regions (paired *t*-tests, $p < .05$).

Table 5

Group fMRI results for the stimulus-based > rule-based t-contrast.

Region	Side	Peak voxel MNI coordinates x y z (mm)	Cluster size	Z score
Caudate	left	-10 12 8	111	5.05
	right	12 16 0	151	5.62
Putamen	left	-16 10 -4	31	4.78
	right	14 8 -8	7	3.74
Insula	left	-30 22 4	62	5.40
	right	34 22 2	51	5.48

6.3.2 Model-based fMRI

6.3.2.1 Statistical analyses

The model-based hypotheses were investigated employing a similar approach to the one used in the model-free statistical analyses. For the first-level analyses, the predictors of interest remained: rule-based choice, rule-based feedback, stimulus-based choice and stimulus-based feedback. However, this time, two separate first-level models were run. In the first model, all predictors of interest were modulated by the trial-wise estimates from the rule network (explained below). In the second model, all predictors of interest were modulated by the trial-wise estimates from the configural memory network (explained below). Within each model, the feedback was additionally modulated by time, which was implemented as a regressor containing the corresponding trial-number. This modulation was introduced to capture variation due to

learning effects (i.e. the fact that the more learning progressed the less participants relied on the feedback).

As far as the estimates in the rule network are concerned, the variable of interest in this analysis was the evidence ratio computed by this network. As explained in **Chapter 5.2,** within the rule network, trial-wise evidence ratios are computed for each stimulus dimension. These ratios are subsequently weighted by the attention paid to each dimension and by the extent to which each dimension was contextually modulated. An average prediction is computed by taking the mean of the weighted ratios of all dimensions. The sign of these predictions signals whether the evidence was for or against the current rule. The absolute value of these predictions indicates the strength of the evidence for the current rule. Since the current analysis focused on explaining rule-learning and not rule correctness, the absolute values were used. This variable is subsequently referred to as *rule prediction*.

With respect to the configural memory, the variable of interest in this analysis was also the evidence ratio computed by this network. As described in **Chapter 5.2,** the memory network computes stimulus-wise evidence ratios. As in the rule network, the sign of this ratios, indicates whether the evidence was for or against a stimulus belonging to a certain category, while the absolute value indicates the strength of the evidence. Since the current analysis focused on memorization per se and not its correctness, the absolute values were taken. This variable is subsequently referred to as *memory prediction.*

Owing to the two-stimulus display, the current adaptation of CAL computes the above-described variables for both the left and the right stimulus. The current analyses included only the predictions corresponding to the chosen stimulus. If a participant failed to respond, the specific trial was discarded. It has to be mentioned that as in the model-free analyses, the model-based analyses concentrated on the choice part of each task. Nevertheless, given that each specific parametric modulator relies on input from the feedback, the variables of interest were also included as parametric modulators for the feedback regressors to account for the feedback-related variance.

The second-level analyses were carried out in the same manner as the model-free second-level analyses with the only differences being that two

separate analyses were run corresponding to the two first-level models, and that at the second level the subject-specific beta images were the rule-based and stimulus based parametric modulators.

6.3.2.2 Results

The hypothesis concerning the higher correlation between the rule network estimates and activity in the prefrontal and hippocampal region in the rule-based task than in the stimulus-based task was assessed using the model with *rule prediction* as parametric modulator. Specifically, the rule prediction parametric modulator corresponding to the choice period in the rule-based task was compared against the rule prediction parametric modulator corresponding to the choice period in the stimulus-based task. Regarding the expected regions of interest, the analysis revealed a correlation between hippocampal activity and rule prediction (whole brain $p < .01$, uncorrected), but the activity did not survive small volume correction. In addition, a significant correlation between rule prediction and vmPFC activity (**Figure 22A**) was found (voxel $_{x, y, z}$ = 2, 36, 0, Z = 3.85, p_{FWE} = 0.01). The first-level beta estimates corresponding to this voxel are displayed in **Figure 22B**. Beta values indicate that the rule-based task was characterized by a weak increase vmPFC activity associated with prediction, while the stimulus-based task was characterized by a decrease in vmPFC associated with rule prediction (comparison significant at $p < .001$). Further exploratory analyses revealed a cluster in the posterior division of the left cingulate gyrus whose peak voxels survived small volume correction within the corresponding anatomical mask (cluster size = 25, peak voxel $_{x, y, z}$ = -8, -42, 38, Z = 4.85, $p_{FWE} < .001$). The beta estimates within these regions exhibited the same pattern as those within vmPFC (also significant at $p < .001$).

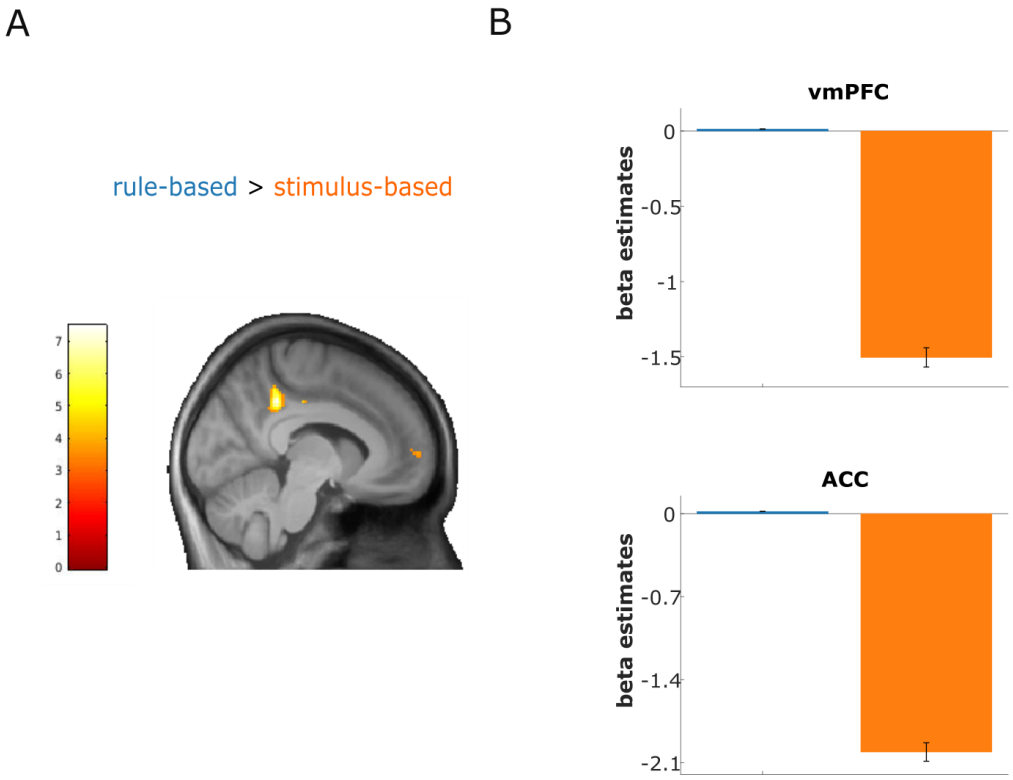


Figure 22. *Model-based fMRI rule prediction results (N = 27).* ***A.*** *Second-level contrast rule-based > stimulus-based. Brighter regions indicate higher t-values. Results thresholded at p < .001 for visualization purposes. Both statistical maps are overlaid onto the group anatomical image (mean T1).* ***B.*** *First-level beta estimates for the peak voxel in the respective region of interest. Blue bars indicate mean betas for the rule prediction parametric modulator in the rule-based task. Orange bars indicate mean betas for the rule prediction parametric modulator in the stimulus-based task. Error bars indicate standard errors.*

The hypothesis posing a higher correlation between configural memory estimates and activity in the striatal regions in the stimulus-based task than in the rule-based task was tested using the model with *memory prediction* as a parametric modulator. The contrast between memory prediction during the stimulus-based task and memory prediction in the rule-based task was investigated. This contrast did not reveal any activity within the caudate, putamen or nucleus accumbens. Further exploration indicated a cluster of activation in the right supplementary motor area, but the constituent voxels did not survive small volume correction in the corresponding anatomical mask. Further exploratory analyses (**Figure 23A**) indicated memory prediction correlates within the right precentral gyrus (peak voxel $_{x, y, z}$ = 44, 4, 30, Z = 4.09, p_{FWE} = 0.01) and left middle frontal gyrus (cluster size = 106, peak voxel $_{x, y, z}$ = -

50, 22, 34, Z = 4.73, p_{FWE} = 0.001). The beta estimates in both regions (**Figure 23B**) suggested an increase in the stimulus-based task associated with the memory prediction, and a decrease in the rule-based task (paired *t*-test within each anatomical region significant at $p < .001$).

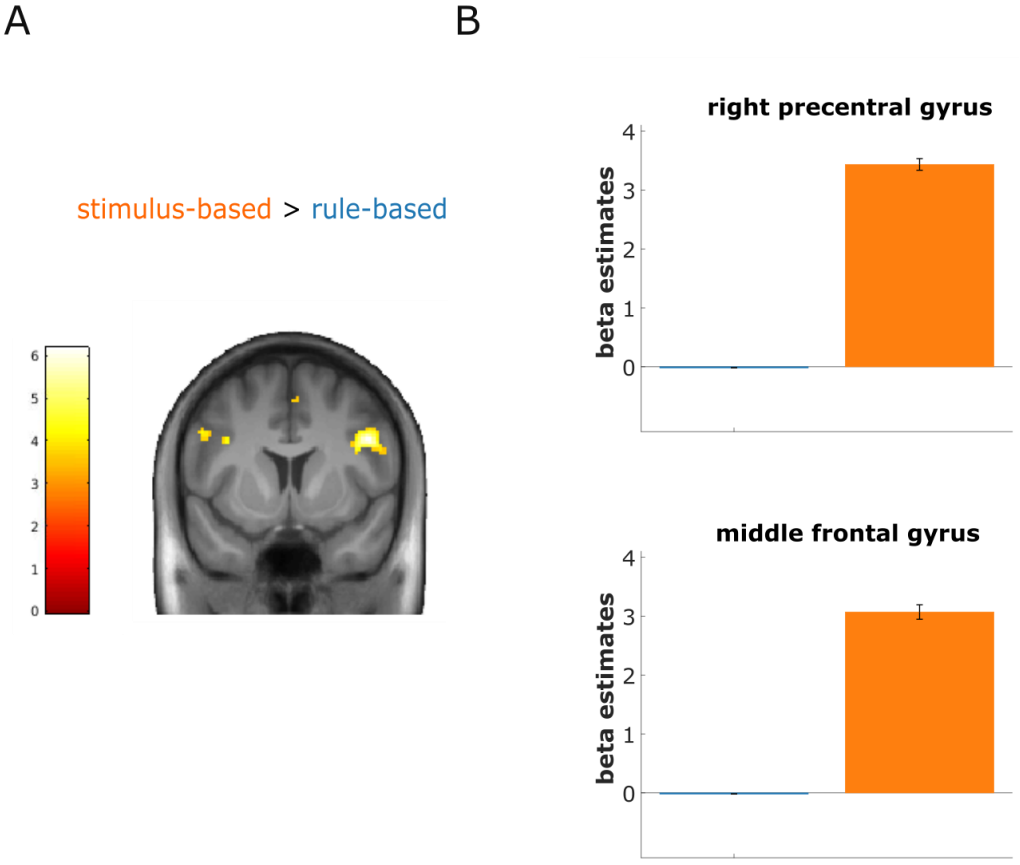


Figure 23. *Model-based fMRI memory prediction results. (N = 27).* ***A.*** *Second-level contrast stimulus-based > rule-based. Brighter regions indicate higher t-values. Results thresholded at p < .001 for visualization purposes. Both statistical maps are overlaid onto the group anatomical image (mean T1).* ***B.*** *First-level beta estimates for the peak voxel in the respective region of interest. Blue bars indicate mean betas for the memory prediction parametric modulator in the rule-based task. Orange bars indicate mean betas for the memory prediction parametric modulator in the stimulus-based task. Error bars indicate standard errors.*

6.4. Discussion

This chapter investigated declarative and procedural categorization by using a new adaptation of Shepard's type II and type VI problems. The two learning types were first addressed using model-free fMRI analyses which revealed more hippocampal and vmPFC involvement in the rule-based task than in the stimulus-based task. By contrast, the stimulus-based task engaged the

ventral striatum and insula more strongly than the rule-based task. The role of these regions was examined further with model-based fMRI analyses using latent variables extracted from the CAL model. CAL was particularly suitable for the current analyses since it offers trial-wise estimates of rule-learning (referred to as rule prediction) and memorization (referred to as memory prediction). As expected, there was higher rule prediction vmPFC activity in the rule-based task as opposed to the stimulus-based task. Although there were indications of hippocampal recruitment, this region did not survive correction for multiple comparisons. Instead, rule prediction activity was found in ACC, whose pattern of activity mirrored the one in the vmPFC. No neural correlates of memory prediction were found in the striatal regions. However, memory prediction related activity was present in the right precentral gyrus and middle frontal gyrus.

One overarching goal of the model-free and model-based analyses was to shed light into the discussion on whether the two learning types involve two distinct neural systems. As far as the model-free analyses are concerned, the second-level results seemingly provide support for the two systems theory where the declarative learning recruits hippocampus and vmPFC and the procedural one recruits the ventral striatum. Nonetheless, the negative first-level beta estimates associated with contrasting the rule-based against the stimulus-based activity should be regarded with care. In both vmPFC and hippocampus, the mean beta-estimates decreased with respect to the model implicit baseline, with the stimulus-based task estimates being significantly lower than the rule-based task estimates. As far as the hippocampus is concerned, previous work has found a general decrease in hippocampal activity during categorization irrespective of the category type (Seger et al., 2011). However, the negative beta estimates could indicate a continuous hippocampal involvement thorough the task. It has been shown that during periods of rest the medial temporal role activity is considerably high even higher than during active baseline condition, and that this effect has previously reversed the sign of activity during memory tasks (Stark & Squire, 2001). It is a possibility that this is also the case in the current task, and that the same rationale could also extend to the vmPFC

estimates and therefore, in both cases the negative beta-estimates do not indicate deactivation.

Regarding the vmPFC differences, parallels could be drawn to the mechanism of neural compression described by Mack et al. (2020). It can be speculated, that in both conditions vmPFC performs information reduction, and the fact that the vmPFC estimates were higher in the rule-based task as opposed to the stimulus-based task is directly correlated with the fact that more information has to be compressed in the rule-based task as opposed to the stimulus-based task.

Concerning the hippocampal results, in light of the well-known general role of hippocampus in encoding and retrieving of objects from memory (Squire, 2004) and the recent findings of its correlation with recognition strength in rule-plus-exception categorization (Davis et al., 2012a) it was not surprising that the region was involved in both tasks. The higher recruitment in the rule-based task compared to the stimulus-based task could reflect that in addition to storing and retrieval, the rule-based task also entails a strong coupling between vmPFC and hippocampus during rule-learning (Mack et al., 2016). While vmPFC is compressing the unnecessary information and predicts rules (as found in the present model-based analyses), the hippocampus stores the tested rules and is responsible for dynamically accessing them (Nomura et al., 2007).

The exploratory analyses on which regions are more active in the rule-based task as opposed to the rule-based task also revealed insular activity. No hypotheses were postulated regarding this region, since it is mostly known for its role in emotion regulation (Giuliani et al., 2011; Grecucci et al., 2013; Steward et al., 2016) and in the saliency network (Menon & Uddin, 2010). However, insular recruitment was found in previous categorization work. Although its direct involvement in categorization has not been previously discussed, the findings seem to suggest a role in item recognition (Seger et al., 2000) and correct categorization of deterministic trials in tasks in which not all members of a category have a deterministic association with the category label (Seger et al., 2010). It could be argued that the higher insular activity during the stimulus-based task as opposed to rule-based task is suggestive of a stronger need for

object recognition in the former rather than the latter. Additional interpretations are difficult without further research.

The model-free hypotheses proposing more ventral striatum activity in the stimulus-based task than the rule-based task was confirmed. This fits well with the numerous findings of procedural learning relying more strongly on striatum than declarative learning (Nomura & Reber, 2008; Seger & Miller, 2010). Although significantly smaller than in the stimulus-based, the first-level beta estimates in the rule-based were far from zero. While this could be solely explained by previous work showing that the striatum plays an important role in categorization tasks, regardless of the strategy employed (Seger et al., 2010), alternative explanations are also plausible. The caudate has been found to play an important role in rule switching (Cools et al., 2004), and thus its recruitment could be crucial in the current rule-based task since participants are quite likely switching between multiple rules until finding the correct one. Nonetheless, since the activity in putamen was also far from zero, the more likely explanation is that the positive beta-estimates are reminiscent of prediction errors (O'Doherty et al., 2004). This signal could be higher in the present study than in previous categorization tasks, due to the current tasks giving monetary feedback as opposed to cognitive feedback (Daniel & Pollmann, 2010). It has been previously shown (Haruno & Kawato, 2006) that the caudate and putamen are computing two distinct prediction errors. The putamen computes a stimulus-action-reward association which occurs at stimulus onset. The caudate computes the well-known reward prediction error, the comparison between the outcome and reward at the feedback stage (caudate). Assuming that the underlying cognitive mechanisms closely matched those in CAL's rule network (i.e. evidence ratios are computed for both valuable and not-valuable categories), one could speculate that in the current case both types of prediction errors play a role during the choice stage, and therefore both the caudate and the putamen are simultaneously recruited (perhaps due to the reward prediction error being temporarily revised).

As far as the model-based results are concerned, the rule prediction patterns found in vmPFC and ACC are consistent with previous work in category learning and decision making (Badre et al., 2010; Hartstra et al., 2010; O'Bryan

et al., 2018). In the framework of COVIS, there has been evidence suggesting that vmPFC generates rules and runs hypothesis testing (Ashby & Maddox, 2011; Ashby & Valentin, 2017; Carpenter et al., 2016; Schnyer et al., 2009) and ACC selects the appropriate rule (Maddox & Ashby, 2004).The rule prediction variable computed by CAL encompasses all these processes. Future work could attempt to model these three processes separately. A possibility could be breaking down CAL's rule prediction variable into its constituent parts: uni-dimensional rule, contextual modulation, integration from all three dimensions and correlating each part with neural activity.

With respect to memory prediction, the absence of correlates in the ventral striatum was surprising. The underlying cause could be found in the task or in the mechanisms behind the memory prediction variable. As far as the task type is concerned, it could be the case that the assumption that the stimulus-based task is a suitable substitute for procedural learning might have been erroneous. Within the COVIS framework, previous research has highlighted that a marker for procedural learning is a strong stimulus-response association (Ashby & Valentin, 2017). From its structure, the current stimulus-based task does not allow for such associations to form since a valuable stimulus could be located either left or right of the stimulus. Furthermore, although cumbersome, the stimulus-based task could be solved via explicit verbalizable rules which would directly counteract the assumption of procedural learning. Parallels can be drawn between the current study and the study by Milton and Pothos (2011), which in the context of COVIS used a substitute of the information-integration task that shared the characteristics of the current stimulus-based task. Namely, the authors used a task, referred to as complex task, which could not be solved by either a uni- or a two-dimensional rule and in which a verbalizable solution, although suboptimal and cognitively demanding, was nevertheless possible. This procedural learning substitute did not elicit any activity in the caudate or putamen regions. Therefore, it cannot be completely ruled out that the lack of memory prediction in the ventral striatum could be due to the stimulus-based task not adequately representing the procedural system. However, the strong caudate and putamen recruitment in the model-free analysis make this explanation rather unlikely.

With regard to the memory prediction variable, it should be revised that the parameter λ, which controls the encoding and retrieval strength, has an important role in its calculation. As discussed in **Chapter 5.7**, although the median values of this parameter were higher in the stimulus-based task, many participants also had high λ values in the rule-based task. Moreover, in both tasks, there was a positive correlation between λ values and accuracy. Therefore, it is probable that for a certain sample of participants or for a certain time period during the task, λ was similar between the two tasks. On the other hand, it could be the case that selecting only the predictions for the valuable stimulus might have been inadequate for capturing memory predictions. Future work could attempt to use an average prediction from the left and right stimuli.

Despite its potential limitations, the memory prediction did reveal meaningful activity in the right precentral gyrus and left middle frontal gyrus. These regions were also active in the study by Milton and Pothos (2011) with declarative and procedural learning both recruiting the left middle frontal gyrus, while the right precentral gyrus was unique to the complex (procedural) task. Previous findings indicate that the left middle frontal gyrus is recruited more when the items are more difficult to categorize (near decision boundaries, DeGutis & D'Esposito, 2007). Therefore, it could be argued that the correlation between memory prediction and this region was due to items being harder to categorize in the stimulus-based task than in rule-based task. With regards to the right precentral gyrus, this region has been previously found to play a role in increasing discriminability of stimulus features (Folstein et al., 2013). Its higher role in the stimulus-based task could be due to the fact that memory prediction entails the computation of feature-wise psychological distances between the current stimuli and the previously stored stimuli, for which is crucial that the features are highly discriminable.

While some limitations to this study have already been discussed above, two more are worthy of note. As previously mentioned, the negative beta estimates pose difficulty in interpretation. The use of an active baseline may have prevented this issue. Additionally, an active baseline could have helped in drawing more definite conclusions on the activation unique to each task. Although the interpretation of first-level beta estimates is meaningful, these

estimates are nonetheless computed with respect to the assumed model's implicit baseline. It is encouraged that future fMRI work on Shepard's problems take this aspect into consideration.

Since this was the very first application of CAL to fMRI data, the methodology still has room for improvement. As far as the model itself is concerned, several possible improvements have already been highlighted in **Chapter 5.7**. Concerning the model-based implementation, it could be that applying the rule prediction and memory prediction as parametric modulators of both the choice period and the feedback period might have not been ideal. In future it could be beneficial to break down these variables into estimates unique to stimulus choice and feedback presentation. Future analysis could also greatly benefit from assessing the rule and memory prediction correlates at different stages during the learning process.

The current work sought to contribute to the debate on whether declarative and procedural category learning recruit two distinct neurobiological systems. The evidence from this study tends to favor a single system theory. While both model-free and model-based second-level analyses suggested a system separation, the first-level estimates tend towards a more unitary system. It is definitely not claimed that the two types of learning recruit regions such as the vmPFC, hippocampus and striatum equally. Instead, although declarative and procedural learning might rely on overlapping regions, these regions dynamically adapt to the current learning type by undergoing different computations. Since the current paradigm aimed to dissociate between the two learning types by providing strategy instructions based on the learning type, no statements can be made on whether the two systems are competing as assumed in the COVIS model. Future work could consider using the same tasks without instructions and apply computational modeling to identify the strategies used by each participant. Although the proposition of a unitary system is controversial given the numerous studies suggesting otherwise (summarized in Ashby & Maddox, 2011; Ashby & Valentin, 2017; Wang & Ashby, 2020), the recent study by Carpenter et al. (2016), which based on the criteria of Edmunds et al. (2018) used the least confounded behavioral paradigm, together with the critical review by Wills et al. (2019) forecast a merging of the two systems.

7. Conclusions and Future Directions

This work presented a detailed comparison of rule-based and stimulus-based categorization through in-depth analyses of behavioral, ocular, computational and neural mechanisms. To carry out this comparison, a novel two-way categorization paradigm was designed. The starting point in the paradigm development were the influential Shepard et al. (1961) categorization problems, in particular the type II problem and type VI problem. While the former can be optimally solved using a two-dimensional logical rule, the latter can be optimally solved through memorization. This thesis deviated from the standard format of these problems by implementing the more ecologically valid two-stimulus display and probabilistic monetary feedback. To obtain a cleaner picture of the two learning types, unconfounded by strategy switches, the paradigm was accompanied by instructions on how to optimally perform each problem type. To capture the studied strategies, the current adaptation of type II and type VI problems were referred to as rule-based and stimulus-based categorization.

Both rule-based and stimulus-based categorization were solved at the same rate, making this the first study to find a lack of type II learning advantage with respect to type VI problems. The two tasks showed a striking behavioral distinction with respect to post-learning RT, which decreased half as much in the rule-based task than in the stimulus-based task. An overall distinction in mean accuracy was present, but minimal.

The two tasks were characterized by distinguishable attentional strategies. Learning a rule entailed that participants' attention dynamically changed during the task, and upon learning, more attention was allocated to rule-forming dimensions. By contrast, memorization required little fluctuation in attention throughout the task, with attention being evenly distributed (covertly or overtly) on all stimuli dimensions. The eye-tracking results favored the line of research which proposes that participants start by fixating all dimensions (e.g. ALCOVE) to the line of research which assumes that only one dimension is initially fixated (RULEX). Importantly, the present data corroborated the previously controversial finding that attention to task-irrelevant dimensions does not cease before learning.

A powerful new computational model (CAL) was applied to the behavioral data to enhance understanding of the cognitive processes underlying the two learning strategies. Particularly noteworthy is the fact that CAL succeeded in capturing the unprecedented equal learning speed of the rule-based and stimulus-based tasks. CAL's free parameters unraveled that strong encoding and retrieval abilities are crucial for high performance in both tasks, and that participants generalize more sharply about members of the target (valuable) category than the non-target (not-valuable) category. In addition, participants in the rule-based task displayed narrower generalization gradients, which suggested optimal application of the instructed rule strategies.

Conventional fMRI analyses outlined differential recruitment of candidate brain regions. The rule-based task elicited more vmPFC and hippocampal activity while the stimulus-based task recruited the insula, caudate and putamen more strongly. By applying CAL for the first time to neural data, this thesis managed to show that computing rule predictions requires vmPFC and ACC involvement. On the other hand, memory predictions are associated with frontal regions, namely the right precentral gyrus and left middle frontal gyrus. Parallels could be drawn between the present study and the recently introduced concept of neural compression (Mack et al., 2020).

It is strongly believed that the newly developed paradigm will open many exciting research avenues. An important matter for future work is the implementation of a generalization phase in both learning tasks. Generalization is a crucial part of categorization (Seger & Peterson, 2013), and thereby the understanding of these two types will not be complete without understanding how new items are approached. Furthermore, much is left to explore concerning the current paradigm. For example, one could investigate the effect of different instruction types or lack thereof on the current tasks. Particularly beneficial for the literature would be assessing the effect of instructions encouraging rule-like and memorization strategies on the stimulus-based task. It has to be highlighted that this thesis concentrated mainly on the processes preceding the categorization decision. Future studies could attempt a detailed examination of attentional and neural mechanisms during feedback, by examining effects of negative and positive feedback on the learning strategies.

Work is being currently done to find new cognitively sound approaches to better address the probabilistic feedback within the CAL framework. It is strongly believed that these improvements will greatly benefit from the model-based fMRI approaches. An unaddressed potential of CAL lies in the model's attentional components. In future, these estimates will be analyzed in great detail and compared against current and future eye-tracking data sets. This has the potential of not only further improving model fit, but also achieving the most detailed description of type II and type VI problems. Although ambitious, it can be posed that, in future, CAL's generalization gradients could be a new modality to examine whether a rule-based strategy was used or not.

An exciting possibility is also the study of strategy switch within the two tasks. Ashby and Maddox (2011) argue that interactions between procedural and declarative learning can be best investigated by employing paradigms in which subjects have to alternate between the two. In this vein, it could be possible to merge the rule-based task and stimulus-based task into one paradigm.

To conclude, in light of all the above findings one could propose that with adequate instructions on the optimal strategies, rule-based and stimulus-based categorization can be learned equally fast. Regardless of the categorization problem, good performance is associated with strong encoding and retrieval abilities. Stimulus-based categorization will nevertheless remain more time consuming than rule-based categorization due to participants needing to allocate a considerable amount of attention to all stimuli dimensions, as opposed to just the relevant ones. Both categorization types recruit a complex network consisting of vmPFC, dorsal PFC, ACC, hippocampus and ventral striatum which adapt their activation and interactions accordingly to the specific problem type. Hence, rule-based and stimulus-based categorization are dissociable category learning systems.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429. https://doi.org/10.1037/0033-295X.98.3.409

Arbel, Y., Feeley, E., & He, X. (2020). The Effect of Feedback on Attention Allocation in Category Learning: An Eye Tracking Study. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.559334

Ardia, D., Mullen, K., Peterson, B. G., Ulrich, J., Boudt, K., & Mullen, M. K. (2015). *DEoptim: Differential evolution in R, R package version, 2.2-4*.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481. https://doi.org/10.1037/0033-295X.105.3.442

Ashby, F. G., Ell, S. W., & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition*, *31*(7), 1114–1125. https://doi.org/10.3758/BF03196132

Ashby, F. G., & Maddox, W. T. (2005). Human Category Learning. *Annual Review of Psychology*, *56*(1), 149–178. https://doi.org/10.1146/annurev.psych.56.091103.070217

Ashby, F. G., & Maddox, W. T. (2011). Human Category Learning 2.0. *Annals of the New York Academy of Sciences*, *1224*, 147–161. https://doi.org/10.1111/j.1749-6632.2010.05874.x

Ashby, F. G., & Valentin, V. V. (2017). Chapter 7 - Multiple Systems of Perceptual Category Learning: Theory and Cognitive Tests. In H. Cohen & C.

Lefebvre (Eds.), *Handbook of Categorization in Cognitive Science (Second Edition)* (pp. 157–188). Elsevier. https://doi.org/10.1016/B978-0-08-101107-2.00007-5

Badre, D., Kayser, A. S., & D'Esposito, M. (2010). Frontal Cortex and the Discovery of Abstract Action Rules. *Neuron*, *66*(2), 315–326. https://doi.org/10.1016/j.neuron.2010.03.025

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3). https://doi.org/10.1016/j.jml.2012.11.001

Barthelme, S. (2019). *eyelinker: Import ASC Files from EyeLink Eye Trackers* (0.2.0) [Computer software]. https://CRAN.R-project.org/package=eyelinker

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *Lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7.*

Bayer, J., Rusch, T., Zhang, L., Gläscher, J., & Sommer, T. (2020). Dose-dependent effects of estrogen on prediction error related neural activity in the nucleus accumbens of healthy young women. *Psychopharmacology*, *237*(3), 745–755. https://doi.org/10.1007/s00213-019-05409-7

Best, C. A., Yim, H., & Sloutsky, V. M. (2013). The cost of selective attention in category learning: Developmental differences between adults and infants. *Journal of Experimental Child Psychology*, *116*(2), 105–119. https://doi.org/10.1016/j.jecp.2013.05.002

Blair, M. R., Watson, M. R., Walshe, R. C., & Maj, F. (2009). Extremely selective attention: Eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1196–1206. https://doi.org/10.1037/a0016272

Cantwell, G., Riesenhuber, M., Roeder, J. L., & Ashby, F. G. (2017). Perceptual category learning and visual processing: An exercise in computational cognitive neuroscience. *Neural Networks*, *89*, 31–38. https://doi.org/10.1016/j.neunet.2017.02.010

Carpenter, K. L., Wills, A. J., Benattayallah, A., & Milton, F. (2016). A Comparison of the neural correlates that underlie rule-based and information-integration category learning. *Human Brain Mapping*, *37*(10), 3557–3574. https://doi.org/10.1002/hbm.23259

Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1699–1719. https://doi.org/10.1037/xlm0000406

Casale, M. B., Roeder, J. L., & Ashby, F. G. (2012). Analogical transfer in perceptual categorization. *Memory & Cognition*, *40*(3), 434–449. https://doi.org/10.3758/s13421-011-0154-4

Clithero, J. A., & Rangel, A. (2014). Informatic parcellation of the network involved in the computation of subjective value. *Social Cognitive and Affective Neuroscience*, *9*(9), 1289–1302. https://doi.org/10.1093/scan/nst106

Cools, R., Clark, L., & Robbins, T. W. (2004). Differential Responses in Human Striatum and Prefrontal Cortex to Changes in Object and Rule Relevance. *Journal of Neuroscience*, *24*(5), 1129–1135. https://doi.org/10.1523/JNEUROSCI.4312-03.2004

Davis, T., Love, B. C., & Preston, A. R. (2012a). Learning the Exception to the Rule: Model-Based fMRI Reveals Specialized Representations for Surprising Category Members. *Cerebral Cortex*, *22*(2), 260–273. https://doi.org/10.1093/cercor/bhr036

Davis, T., Love, B. C., & Preston, A. R. (2012b). Striatal and hippocampal entropy and recognition signals in category learning: Simultaneous processes revealed by model-based fMRI. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*(4), 821–839. https://doi.org/10.1037/a0027865

DeGutis, J., & D'Esposito, M. (2007). Distinct mechanisms in visual category learning. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(3), 251–259. https://doi.org/10.3758/CABN.7.3.251

Deng, W. (Sophia), & Sloutsky, V. M. (2016). Selective Attention, Diffused Attention, and the Development of Categorization. *Cognitive Psychology*, *91*, 24–62. https://doi.org/10.1016/j.cogpsych.2016.09.002

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, *22*(6), 1075–1081. https://doi.org/10.1016/j.conb.2012.08.003

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2018). Due Process in Dual Process: Model-Recovery Simulations of Decision-Bound Strategy Analysis in

Category Learning. *Cognitive Science*, *42*(S3), 833–860. https://doi.org/10.1111/cogs.12607

Erickson, M., & Kruschke, J. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology. General*. https://doi.org/10.1037/0096-3445.127.2.107

Filoteo, J., Maddox, W., & Davis, J. (2001). A possible role of the striatum in linear and nonlinear category learning: Evidence from patients with Huntington's disease. *Behavioral Neuroscience*, *115*, 786–798. https://doi.org/10.1037/0735-7044.115.4.786

Filoteo, J. V., Salmon, D. P., Maddox, W. T., & Song, D. D. (2005). Informationintegration category learning in patients with striatal dysfunction. *Neuropsychology*, 212–222.

Folstein, J. R., Palmeri, T. J., & Gauthier, I. (2013). Category Learning Increases Discriminability of Relevant Object Dimensions in Visual Cortex. *Cerebral Cortex*, *23*(4), 814–823. https://doi.org/10.1093/cercor/bhs067

FSL - FslWiki. (n.d.). Retrieved February 14, 2021, from https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/

Georgioudakis, M., & Plevris, V. (2020). A Comparative Study of Differential Evolution Variants in Constrained Structural Optimization. *Frontiers in Built Environment*, *6*. https://doi.org/10.3389/fbuil.2020.00102

Giuliani, N. R., Drabant, E. M., Bhatnagar, R., & Gross, J. J. (2011). Emotion regulation and brain plasticity: Expressive suppression use predicts anterior insula volume. *NeuroImage*, *58*(1), 10–15. https://doi.org/10.1016/j.neuroimage.2011.06.028

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. *Neuron*, *66*(4), 585–595. https://doi.org/10.1016/j.neuron.2010.04.016

Gläscher, J., Hampton, A. N., & O'Doherty, J. P. (2009). Determining a Role for Ventromedial Prefrontal Cortex in Encoding Action-Based Value Signals During Reward-Related Decision Making. *Cerebral Cortex*, *19*(2), 483–495. https://doi.org/10.1093/cercor/bhn098

Gläscher, J. P., & O'Doherty, J. P. (2010). Model-based approaches to neuroimaging: Combining reinforcement learning theory with fMRI data. *WIREs Cognitive Science*, *1*(4), 501–510. https://doi.org/10.1002/wcs.57

Gluck, M. A., Bower, G. H., & Hee, M. R. (1989, August). A configural-cue network model of animal and human associative learning. *In Proceedings of the Eleventh Annual Conference of the Cognitive Science Society (Vol. 11, Pp. 323-332). Erlbaum Ann Arbor, Michigan. Hillsdale, NJ.*

Gluck, M. A., Glauthier, P. T., & Sutton, R. S. (1992). Adaptation of cue-specific learning rates in network models of human category learning. *In Proceedings of the Fourteenth Annual Meeting of the Cognitive Science Society Bloomington, IN.*

Grecucci, A., Giorgetta, C., Bonini, N., & Sanfey, A. G. (2013). Reappraising social emotions: The role of inferior frontal gyrus, temporo-parietal junction and insula in interpersonal emotion regulation. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00523

Hartstra, E., Oldenburg, J. F. E., Van Leijenhorst, L., Rombouts, S. A. R. B., & Crone, E. A. (2010). Brain regions involved in the learning and application of reward rules in a two-deck gambling task. *Neuropsychologia*, *48*(5), 1438–1446. https://doi.org/10.1016/j.neuropsychologia.2010.01.012

Haruno, M., & Kawato, M. (2006). Different Neural Correlates of Reward Expectation and Reward Expectation Error in the Putamen and Caudate Nucleus During Stimulus-Action-Reward Association Learning. *Journal of Neurophysiology*, *95*(2), 948–959. https://doi.org/10.1152/jn.00382.2005

Hoffman, A. B., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General*, *139*(2), 319–340. https://doi.org/10.1037/a0019042

Kim, S., & Rehder, B. (2011). How prior knowledge affects selective attention during category learning: An eyetracking study. *Memory & Cognition*, *39*(4), 649–665. https://doi.org/10.3758/s13421-010-0050-3

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44. https://doi.org/10.1037/0033-295X.99.1.22

KRUSCHKE, J. K. (1993). Human Category Learning: Implications for Backpropagation Models. *Connection Science*, *5*(1), 3–36. https://doi.org/10.1080/09540099308915683

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, *45*, 812–863.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, *14*(4), 560–576. https://doi.org/10.3758/BF03196806

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: Revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 552–572. https://doi.org/10.1037/a0029178

Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 720–738. https://doi.org/10.1037/a0022639

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Sciences*, *4*(1), 6–14. https://doi.org/10.1016/S1364-6613(99)01418-7

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review*, *9*(4), 829–835. https://doi.org/10.3758/BF03196342

Love, B. C., & Markman, A. B. (2003). The nonindependence of stimulus properties in human category learning. *Memory & Cognition*, *31*(5), 790–799. https://doi.org/10.3758/BF03196117

Love, B., Medin, D., & Gureckis, T. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309–332. https://doi.org/10.1037/0033-295X.111.2.309

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, *113*(46), 13203–13208. https://doi.org/10.1073/pnas.1614048113

Mack, M. L., Love, B. C., & Preston, A. R. (2018). Building concepts one episode at a time: The hippocampus and concept formation. *Neuroscience Letters*, *680*, 31–38. https://doi.org/10.1016/j.neulet.2017.07.061

Mack, M. L., Preston, A. R., & Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nature Communications*, *11*(1), 46. https://doi.org/10.1038/s41467-019-13930-8

Maddox, W. T., & Ashby, F. G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioural Processes*, *66*(3), 309–332. https://doi.org/10.1016/j.beproc.2004.03.011

Matsuka, T., & Corter, J. E. (2008). Observed attention allocation processes in category learning. *Quarterly Journal of Experimental Psychology*, *61*(7), 1067–1097. https://doi.org/10.1080/17470210701438194

McColeman, C. M., Barnes, J. I., Chen, L., Meier, K. M., Walshe, R. C., & Blair, M. R. (2014). Learning-Induced Changes in Attentional Allocation during Categorization: A Sizable Catalog of Attention Change as Measured by

Eye Movements. *PLOS ONE*, *9*(1), e83302. https://doi.org/10.1371/journal.pone.0083302

Menon, V., & Uddin, L. Q. (2010). *Saliency, switching, attention and control: A network model of insula function*.

Milton, F., & Pothos, E. M. (2011). Category structure and the two learning systems of COVIS. *European Journal of Neuroscience*, *34*(8), 1326–1336. https://doi.org/10.1111/j.1460-9568.2011.07847.x

MP100 Starter Systems | BIOPAC. (n.d.). Retrieved February 1, 2021, from https://www.biopac.com/product-category/mp100-starter-systems/

Nomura, E. M., & Reber, P. J. (2008). A review of medial temporal lobe and caudate contributions to visual category learning. *Neuroscience & Biobehavioral Reviews*, *32*(2), 279–291. https://doi.org/10.1016/j.neubiorev.2007.07.006

Nomura, E., Maddox, W., Filoteo, J., Ing, A., Gitelman, D., Parrish, T., Mesulam, M.-M., & Reber, P. (2007). Neural Correlates of Rule-Based and Information-Integration Visual Category Learning. *Cerebral Cortex*, *17*(1), 37–43. https://doi.org/10.1093/cercor/bhj122

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., Mckinley, S. C., & Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, *22*(3), 352–369. https://doi.org/10.3758/BF03200862

Nosofsky, R. M., & Kruschke, J. K. (2002). Single-system models and interference in category learning: Commentary on Waldron and Ashby

(2001). *Psychonomic Bulletin & Review*, *9*(1), 169–174. https://doi.org/10.3758/BF03196274

Nosofsky, R. M., & Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review*, *5*(3), 345–369. https://doi.org/10.3758/BF03208813

O'Bryan, S. R., Walden, E., Serra, M. J., & Davis, T. (2018). Rule activation and ventromedial prefrontal engagement support accurate stopping in self-paced learning. *NeuroImage*, *172*, 415–426. https://doi.org/10.1016/j.neuroimage.2018.01.084

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. *Science*, *304*(5669), 452–454. https://doi.org/10.1126/science.1094285

Pape, A. D., & Kurtz, K. J. (2013). Evaluating case-based decision theory: Predicting empirical patterns of human classification learning. *Games and Economic Behavior*, *82*, 52–65. https://doi.org/10.1016/j.geb.2013.06.010

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042–1045. https://doi.org/10.1038/nature05051

Pickering, A. D. (1997). New Approaches to the Study of Amnesic Patients: What Can a Neurofunctional Philosophy and Neural Network Methods Offer? *Memory*, *5*(1–2), 255–300. https://doi.org/10.1080/741941146

Pothos, E. M. (2005). The rules versus similarity distinction. *Behavioral and Brain Sciences*, *28*(1), 1–14. https://doi.org/10.1017/S0140525X05000014

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*(1), 1–41. https://doi.org/10.1016/j.cogpsych.2004.11.001

Schlegelmilch, R., Wills, A., & Helversen, B. von. (2021). *A Cognitive Category-Learning Model of Rule Abstraction, Attention Learning, and Contextual Modulation*. PsyArXiv. https://doi.org/10.31234/osf.io/4jukw

Schnyer, D. M., Maddox, W. T., Ell, S., Davis, S., Pacheco, J., & Verfaellie, M. (2009). Prefrontal contributions to rule-based and information-integration category learning. *Neuropsychologia*, *47*(13), 2995–3006. https://doi.org/10.1016/j.neuropsychologia.2009.07.011

Seger, C. A., Dennison, C. S., Lopez-Paniagua, D., Peterson, E. J., & Roark, A. A. (2011). Dissociating hippocampal and basal ganglia contributions to category learning using stimulus novelty and subjective judgments. *NeuroImage*, *55*(4), 1739–1753. https://doi.org/10.1016/j.neuroimage.2011.01.026

Seger, C. A., & Miller, E. K. (2010). Category Learning in the Brain. *Annual Review of Neuroscience*, *33*(1), 203–219. https://doi.org/10.1146/annurev.neuro.051508.135546

Seger, C. A., & Peterson, E. J. (2013). Categorization = Decision Making + Generalization. *Neuroscience and Biobehavioral Reviews*, *37*(7), 1187–1200. https://doi.org/10.1016/j.neubiorev.2013.03.015

Seger, C. A., Peterson, E. J., Cincotta, C. M., Lopez-Paniagua, D., & Anderson, C. W. (2010). Dissociating the contributions of independent corticostriatal systems to visual categorization learning through the use of reinforcement learning modeling and Granger causality modeling. *NeuroImage*, *50*(2), 644–656. https://doi.org/10.1016/j.neuroimage.2009.11.083

Seger, C. A., Prabhakaran, V., Poldrack, R. A., & Gabrieli, J. D. E. (2000). Neural activity differs between explicit and implicit learning of artificial grammar strings: An fMRI study. *Psychobiology*, *28*(3), 283–292. https://doi.org/10.3758/BF03331987

Sewell, D. K., Warren, H. A., Rosenblatt, D., Bennett, D., Lyons, M., & Bode, S. (2018). Feedback Discounting in Probabilistic Categorization: Converging Evidence from EEG and Cognitive Modeling. *Computational Brain & Behavior*, *1*(2), 165–183. https://doi.org/10.1007/s42113-018-0012-6

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42. https://doi.org/10.1037/h0093825

Singmann, H., Bolker, B., Westfall, J., Aust, F., Ben-Shachar, M. S., Højsgaard, S., Fox, J., Lawrence, M. A., Mertens, U., Love, J., Lenth, R., & Christensen, R. H. B. (2021). *afex: Analysis of Factorial Experiments* (0.28-1) [Computer software]. https://CRAN.R-project.org/package=afex

Smith, A. C., Frank, L. M., Wirth, S., Yanike, M., Hu, D., Kubota, Y., Graybiel, A. M., Suzuki, W. A., & Brown, E. N. (2004). Dynamic Analysis of Learning

in Behavioral Experiments. *Journal of Neuroscience*, *24*(2), 447–461. https://doi.org/10.1523/JNEUROSCI.2908-03.2004

Smith, J. D., Boomer, J., Zakrzewski, A. C., Roeder, J. L., Church, B. A., & Ashby, F. G. (2014). Deferred Feedback Sharply Dissociates Implicit and Explicit Category Learning. *Psychological Science*, *25*(2), 447–457. https://doi.org/10.1177/0956797613509112

Smith, J. D., & Church, B. A. (2018). Dissociable learning processes in comparative psychology. *Psychonomic Bulletin & Review*, *25*(5), 1565–1584. https://doi.org/10.3758/s13423-017-1353-1

Smith, J. D., Minda, J. P., & Washburn, D. A. (2004). Category Learning in Rhesus Monkeys: A Study of the Shepard, Hovland, and Jenkins (1961) Tasks. *Journal of Experimental Psychology: General*, *133*(3), 398–414. https://doi.org/10.1037/0096-3445.133.3.398

SPM - Documentation. (n.d.). Retrieved February 3, 2021, from https://www.fil.ion.ucl.ac.uk/spm/doc/

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, *82*(3), 171–177. https://doi.org/10.1016/j.nlm.2004.06.005

Stanford, T. R., Shankar, S., Massoglia, D. P., Costello, M. G., & Salinas, E. (2010). Perceptual decision making in less than 30 milliseconds. *Nature Neuroscience*, *13*(3), 379–385. https://doi.org/10.1038/nn.2485

Stanton, R. D., & Nosofsky, R. M. (2007). Feedback interference and dissociations of classification: Evidence against the multiple-learning-

systems hypothesis. *Memory & Cognition*, *35*(7), 1747–1758. https://doi.org/10.3758/BF03193507

Stanton, R. D., & Nosofsky, R. M. (2013). Category number impacts rule-based and information-integration category learning: A reassessment of evidence for dissociable category-learning systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1174–1191. https://doi.org/10.1037/a0031670

Stark, C. E. L., & Squire, L. R. (2001). When zero is not zero: The problem of ambiguous baseline conditions in fMRI. *Proceedings of the National Academy of Sciences*, *98*(22), 12760–12766. https://doi.org/10.1073/pnas.221462998

Steward, T., Picó-Pérez, M., Mata, F., Martínez-Zalacaín, I., Cano, M., Contreras-Rodríguez, O., Fernández-Aranda, F., Yucel, M., Soriano-Mas, C., & Verdejo-García, A. (2016). Emotion Regulation and Excess Weight: Impaired Affective Processing Characterized by Dysfunctional Insula Activation and Connectivity. *PLOS ONE*, *11*(3), e0152150. https://doi.org/10.1371/journal.pone.0152150

Team, R. C. (2013). *R: A language and environment for statistical computing*.

van Renswoude, D. R., Raijmakers, M. E. J., Koornneef, A., Johnson, S. P., Hunnius, S., & Visser, I. (2018). Gazepath: An eye-tracking analysis tool that accounts for individual differences and data quality. *Behavior Research Methods*, *50*(2), 834–852. https://doi.org/10.3758/s13428-017-0909-3

Vigo, R., Zeigler, D. E., & Halsey, P. A. (2013). Gaze and informativeness during category learning: Evidence for an inverse relation. *Visual Cognition*, *21*(4), 446–476. https://doi.org/10.1080/13506285.2013.800931

Waldschmidt, J. G., & Ashby, F. G. (2011). Cortical and striatal contributions to automaticity in information-integration categorization. *NeuroImage*, *56*(3), 1791–1802. https://doi.org/10.1016/j.neuroimage.2011.02.011

Wang, Y.-W., & Ashby, F. G. (2020). A role for the medial temporal lobes in category learning. *Learning & Memory*, *27*(10), 441–450. https://doi.org/10.1101/lm.051995.120

Watson, M. R., & Blair, M. R. (2008). *Attentional allocation during feedback: Eyetracking adventures on the other side of the response*.

Wills, A. J., Edmunds, C. E. R., Le Pelley, M. E., Milton, F., Newell, B. R., Dwyer, D. M., & Shanks, D. R. (2019). Dissociable learning processes, associative theory, and testimonial reviews: A comment on Smith and Church (2018). *Psychonomic Bulletin & Review*, *26*(6), 1988–1993. https://doi.org/10.3758/s13423-019-01644-3

Wilson, R. C., & Niv, Y. (2015). Is Model Fitting Necessary for Model-Based fMRI? *PLOS Computational Biology*, *11*(6), e1004237. https://doi.org/10.1371/journal.pcbi.1004237

Zaki, S. R., & Salmi, I. L. (2019). Sequence as context in category learning: An eyetracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(11), 1942–1954. https://doi.org/10.1037/xlm0000693

List of Figures

List of Tables

Curriculum Vitae

Lebenslauf wurde aus datenschutzrechtlichen Gründen entfernt.

Eidesstattliche Versicherung

Ich versichere ausdrücklich, dass ich die Arbeit selbständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die aus den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen einzeln nach Ausgabe (Auflage und Jahr des Erscheinens), Band und Seite des benutzten Werkes kenntlich gemacht habe.

Ferner versichere ich, dass ich die Dissertation bisher nicht einem Fachvertreter an einer anderen Hochschule zur Überprüfung vorgelegt oder mich anderweitig um Zulassung zur Promotion beworben habe.

Ich erkläre mich einverstanden, dass meine Dissertation vom Dekanat der Medizinischen Fakultät mit einer gängigen Software zur Erkennung von Plagiaten überprüft werden kann.

Unterschrift: .....