Universität Basel
Fakultät für Psychologie

Similarity-based and abstraction-based processes in judgments from multiple information sources

Inauguraldissertation zur Erlangung der Würde eines Doktors der Philosophie vorgelegt der Fakultät für Psychologie der Universität Basel von

Rebecca Albrecht-Dietsch

aus Heidelberg

Basel, 2020

Universität Basel

Fakultät für Psychologie

Genehmigt von der Fakultät für Psychologie auf Antrag von

Prof. Dr. Jörg Rieskamp

Prof. Dr. Bettina von Helversen

Datum des Doktoratsexamen:

Prof. Dr. Jens Gaab

Universität Basel

Fakultät für Psychologie

Erklärung zur wissenschaftlichen Lauterkeit

Ich erkläre hiermit, dass die vorliegende Arbeit ohne die Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel selbstständig verfasst habe. Zu Hilfe genommene Quellen sind als solche gekennzeichnet. Die veröffentlichten oder zur Veröffentlichung in Zeitschriften eingereichten Manuskripte wurden in Zusammenarbeit mit den Koautoren erstellt und von keinem der Beteiligten an anderer Stelle publiziert, zur Publikation eingereicht, oder einer anderen Prüfungsbehörde als Qualifikationsarbeit vorgelegt.* Es handelt sich dabei um folgende Manuskripte:

- Albrecht, R., Hoffmann, J. A., Pleskac, T. J., Rieskamp, J., & von Helversen, B. (2020a). Competitive retrieval strategy causes multimodal response distributions in multiple-cue judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(6), 1064-1090.
  Primarily responsible for development of the cognitive model, experimental paradigm and design, data collection and analyses, and writing of the paper.
- Albrecht**, R., Jenny**, M. A., Nilsson, H., & Rieskamp, J. (2020b). The Similarity-Updating Model of Probability Judgment and Belief Revision. Manuscript under revision.
  * A previous version of this manuscript was submitted by Mirjam A. Jenny as part of her dissertation in Psychology at the University of Basel (2013). The version submitted here differs significantly from the previous one as it has a new focus, a changed cognitive model, a new competitor model and model comparison with changed methodology, an additional experimental study, and additional literature research and description. I am a part of this research project since 2017 and I primarily developed and implemented these new aspects.
  ** These authors contributed equally to this work.
- Hoffmann, J.A., Albrecht, R., von Helversen, B. (2020c). Coordinating several mental strategies requires integration: Evidence from human judgment. Manuscript submitted for publication.
  Jointly responsible for designing the research and writing the manuscript. Primarily responsible for implementing the experiments.

Basel, 19.08.2020

Rebecca Albrecht-Dietsch

Acknowledgements

First and foremost, I would like to thank my advisors Bettina von Helversen, Jörg Rieskamp, and Janina Hoffman. They always pushed me forward and kept me on the right track while still giving me enough space to explore my own ideas. I want to especially thank Bettina von Helervsen who always and immediately had an open ear for all my questions and problems. It was a great experience working with each of you and I am very grateful for that.

I would like to thank my friends and current and former colleagues from the EconPsych in Basel and the CogSci in Freiburg who have accompanied me through parts of this academic ride by having great discussions, giving good advice, always challenging my thoughts and not letting me settle down to much in comfortable mind-spaces. Especially, I would like to thank Laura Fontanesi for the fun time in the shared office, Mikhail Spektor for always knowing everything and never hesitating to share, Jana Jarecki for giving great feedback, and Rul von Stülpnagel for working with me on projects although I had so little time. Also, I would like thank my student assistant, Florian Seitz, for his great commitment. A special thanks I want to forward to Markus Lohmeyer for endless discussions which shaped my thinking deeply. I also want to thank Jana Jarecki, Mikhail Spektor, Florian Seitz, Janina Hoffman, and Dimitris Katsimpokis for giving me valuable feedback for my dissertation.

A very special thanks go to all my family. My kids Nicolas and Sophie who always brought so much joy and light even to a cloudy day. My husband Daniel who always loved me and supported me unconditionally, sometimes at great personal and professional cost. My mum for always supporting all of my decisions and always putting everything else aside to come to help us. My parents-in-law for creating a kind and loving environment for our family and especially their grand-kids. My dad, his partner Christine, and my brothers Raphael and Robert with his family for showing us another world and giving us asylum from ours from time to time.

Abstract

In this dissertation, I propose that similarity-based processes are an integral part of the judgment process and that they are sequentially integrated and interact with abstraction-based processes in a weighted, additive fashion. This theoretical stance challenges the state-of-the-art in the judgment domain that people shift between similarity-based and abstraction-based processes depending on several factors. As such, this dissertation provides new explanations for how judgments are formed and empirical evidence in support of the theoretical foundations. Specifically, it explains the empirical finding that people are in general able to abstract knowledge about the relationship between objects and associated outcomes and that at the same time similarities to known situations bias judgments, although they are assumed to be not informative. In the first paper, I present an anchoring-and-adjustment model which assumes that a judgment is the result of recalling one known instance from memory that is then adjusted based on abstracted knowledge. This model explains how both types of processes interact without the additional need to assume that people actively switch between different processes. In the second paper, I present an anchoring-and-adjustment model that assumes that repeated probability judgments are the result of weighing and adding single probabilities that are the result of a similarity-based process. The result of the weighing-and-adding process is adjusted depending on the similarity between previous judgments and the current context. This model explains how non-informative information influences judgments. In the third paper, I present a general approach designed to understand how people learn to coordinate and apply several different processes. Empirical investigations conclude that people build a weighted average of different cognitive processes and do not switch from one to the other. All in all, the results presented in this dissertation bring important new insights about the combination of different information sources and the coordination of similarity-based and abstraction-based cognitive processes.

Introduction

Similarity is an important concept that lies at the very heart of cognition (Medin, Goldstone, & Gentner, 1993; Tversky, 1977; Attneave, 1950; Hahn, 2014). In general terms, similarity-based theories assume that individuals evaluate a novel problem by comparison to problems, or representations thereof, that have already been solved or are otherwise known. Similarity-based theories are known to researchers in almost every psychological domain: in categorization, and judgment and decision making in the form of prototype (Minda & Smith, 2001; Hampton, 2000) and exemplar theory (Nosofsky, 1988; Medin & Schaffer, 1978; Juslin, Olsson, & Olsson, 2003), in reasoning in the form of case-based reasoning (Hahn & Chater, 1998b), in decisions from experience in the form of instance-based learning (Gonzalez, Lerch, & Lebiere, 2003), in function learning as associative theories (DeLosh, Busemeyer, & McDaniel, 1997), or in the formation of subjective probability judgments as similarity heuristic (Read & Grushka-Cockayne, 2011), and representativeness as prototype similarity/relative likelihood (Nilsson, Olsson, & Juslin, 2005). In categorization, for example, both exemplar theory and prototype theory assume that an object is most likely assigned to the most similar category. The theories differ in that exemplar theory calculates the similarity to all category members (i.e., exemplars), whereas prototype theory calculates the similarity to a category prototype. Similarity has also been studied as an integral part of cognitive functions. For instance, almost all theories of memory retrieval propose that people activate past memories as a function of similarity (for an overview, see Raaijmakers & Shiffrin, 1992; Dougherty, Gettys, & Ogden, 1999; Hintzman, 1984; Lewandowsky, Murdock, et al., 1989; Anderson, 1983).

A novel problem can, however, not only be addressed by comparing it to known problems (similarity-based process). Alternatively, people may use knowledge abstracted from a set of examples (abstraction-based process; Sun, 1995; Juslin, Karlsson, & Olsson, 2008; Hoffmann, von Helversen, & Rieskamp, 2016; DeLosh et al., 1997; Erickson & Kruschke, 1998; Hahn & Chater, 1998a). Abstraction-based theories generally assume that people infer abstract representations

like rules that describe how features of an object or aspects of a situation partly determine a solution. In reasoning, it is often assumed that people use the rules of formal logical systems to derive the solution to a problem (Rips, 1994; Braine & O'Brien, 1998). In categorization, a rule describes which feature values decide over category membership (Erickson & Kruschke, 1998; Nosofsky, Palmeri, & McKinley, 1994), while in judgment the rule is mostly a linear additive function with weights specifying the importance of different features (Juslin, Jones, Olsson, & Winman, 2003; Hoffmann et al., 2016).

While similarity-based and abstraction-based theories have been used to explain findings in the respective fields, both have limitations in explaining certain behaviors. In the domains of judgment and categorization, for example, similarity-based theories have been criticized for not being able to explain human generalization capabilities. The ability to extrapolate beyond the realm of learned examples is a natural consequence of rule application but cannot be explained by standard similarity-based theories like exemplar theory. Thus, analyses of extrapolation behavior have been used to understand which strategy people use in a given task and patterns of extrapolation have often been found (Juslin et al., 2008; Pachur & Olsson, 2012; DeLosh et al., 1997; Juslin, Olsson, & Olsson, 2003; Little & Lewandowsky, 2009; Hoffmann et al., 2016). Yet, it has been shown that patterns predicted by similarity-based processes play a role even in tasks that are known to foster rule-based processing (Hahn, Prat-Sala, Pothos, & Brumby, 2010; von Helversen, Herzog, & Rieskamp, 2014). For example, although it has repeatedly been shown that people apply rules in tasks where a simple, predictive rule can be used to categorize objects (Allen & Brooks, 1991; Regehr & Brooks, 1993), signs of similarity-based processing have still been found in error patterns and response times (Hahn et al., 2010). In a more applied consideration, it has been shown that facial similarities play a role in professional decision situations, for example in the context of parole decisions or hiring of new employees, where decisions should be based on relevant facts alone (von Helversen et al., 2014).

Because both similarity-based and abstraction-based theories of cognition imperfectly explain the empirical findings on their own, researchers in many domains have suggested that people

mix cognitive processes (Sun, 1995; DeLosh et al., 1997; Erickson & Kruschke, 1998; Nosofsky et al., 1994). These mixture theories assume that the results of cognitive processes (or strategies) are integrated to form a task response. On a general level, two conceptually different approaches to integrating cognitive processes can be distinguished. Some theories assume that people shift between different processes to select a fitting one in a given context (strategy shifting; Gigerenzer & Selten, 2002; Evans, 2019; Juslin et al., 2008; Sun, Slusarz, & Terry, 2005; Stanovich & West, 2000; Gershman, Markman, & Otto, 2014). In the domain of judgment, for example, researchers often assume that every decision is only based on one cognitive process, either on abstractions or on similarities (but see Bröder, Gräf, & Kieslich, 2017, for a different approach). People shift between these strategies depending on factors like the type of the environment (Hoffmann et al., 2016; Juslin et al., 2008; Karlsson, Juslin, & Olsson, 2007), the type of task (Pachur & Olsson, 2012; Juslin, Olsson, & Olsson, 2003; Trippas & Pachur, 2019), the type of the feedback (Juslin, Jones, et al., 2003), and the abilities of the decision maker (Hoffmann, von Helversen, & Rieskamp, 2014). Other theories assume that people combine two strategies by, for example, creating a weighted average of the different strategies' responses (strategy blending; Bröder et al., 2017; Söllner, Bröder, Glöckner, & Betsch, 2014; Glöckner & Bröder, 2011; Herzog & Hertwig, 2009). In the domains of function learning and categorization, for example, theories often assume that the results of abstraction-based and similarity-based processes are directly combined into one weighted-additive response (DeLosh et al., 1997; Erickson & Kruschke, 1998). However, it remains largely unclear how people integrate and possibly learn to coordinate the use of different strategies on a general level.

In addition to the combination of different cognitive processes, it is also important to consider how the results of sequentially executed processes are integrated. Additive, sometimes weighted, combinations of sequentially obtained results have been proposed in several different areas. Very generally, theories assuming sequential sampling describe the process of forming a judgment (Nosofsky, 1997) or retrieving an item from memory (Ratcliff, 1978) as the repeated execution of one cognitive processes (for example retrieval of a memory item) with an additive combination

of the results. In belief-revision and belief-updating, a currently held belief might be changed by additional evidence and it has been shown that weighing-and-adding models explain this cognitive process well (Hogarth & Einhorn, 1992). In the judgment domain, sequential, additive combinations of information, e.g. features in abstraction-based processes, have been proposed as the cognitive mechanism underlying rule-based and exemplar-based processes (Juslin et al., 2008). In sum, there is substantial evidence suggesting that cognitive processes are sequentially combined in a weighted, additive fashion.

In this dissertation, I investigate how similarity-based processes are integrated sequentially and interact with abstraction-based processes to form a judgment from multiple features (or cues; introduced in the next section). Specifically, I propose that people judge an object or situation by building a weighted, additive combination of single process responses. First, I introduce an alternative explanation for presumed strategy shifts in the judgment domain by proposing a direct combination of similarity-based memory retrieval and abstracted knowledge about the importance of different features (paper 1). Second, I show how similarity-based subjective probability judgments are combined in a weighted, additive fashion and how the result of this process is biased by similarity (paper 2). Finally, I introduce a general learning model to investigate how similarity-based and abstraction-based cognitive processes are coordinated and present evidence that people combine these processes in the judgment domain by weighing and adding the responses of both processes (paper 3).

Judgments from multiple cues

In this dissertation, I investigate the interaction between similarity-based and abstraction-based processes in the domain of judgments from multiple cues. From a general point of view, judgments from multiple cues describe the estimation of a numeric criterion from a set of cues or features. A cover story often used in judgment and categorization tasks is the assessment of the toxicity of fictitious bugs (Juslin, Olsson, & Olsson, 2003; Pachur & Olsson, 2012; Bröder et al., 2017; Hoffmann et al., 2016). The bugs are characterized by a set of features like the number

of stripes on their bodies or the length of their legs. The numeric criterion that needs to be estimated could be a number that describes some quality of the object (i.e., numeric estimation, "how toxic is this bug") or a probability (i.e., probability judgments, "how likely will this bug kill me"). In the following, I describe research methods and empirical paradigms applied in the fields of numeric estimation (papers 1 and 3) and probability judgments (paper 2) that are relevant for this dissertation.

Numeric estimation. In numeric estimation tasks, researchers try to determine which of the two strategies, similarity-based and abstraction-based, are used by participants. Similarity-based theories, which have been implemented as exemplar models in the judgment domain (Juslin, Olsson, & Olsson, 2003; Hoffmann et al., 2016), assume that the value of a to-be-judged object (or exemplar) is the average of all known objects' values weighted by their similarity to the to-be-judged object. As such, exemplar models are sometimes referred to as following a memory-based strategy (Hoffmann et al., 2014). Revisiting our bug example, the toxicity of one bug would be determined by summing up the toxicity values of all known bugs weighted by their similarity to that bug. A bug, for example, that has a number of stripes and length of legs closely matching the most toxic of creatures would, therefore, be judged as toxic. Abstraction-based theories in the judgment domain are mostly implemented as main-effects linear regression models (Hammond & Stewart, 2001; Cooksey, 1996) and are called rule-based or cue-abstraction models. In these models, the different feature values are summed up and weighted by their relative, subjective importance. In the bug example, this could mean that a person believes the number of stripes to be very predictive of the toxicity value whereas they deem the length of the legs to be unimportant.

In order to determine which of the two strategies people use under certain circumstances to solve numeric estimation tasks, researchers define a functional dependency between the cues and the criterion (called environment). For example, one could define an environment where the toxicity of bugs is calculated as a linear function (linear environment) of the feature values.

During experiments, people receive training where they estimate the criterion based on presented examples and receive feedback. In a subsequent test phase, they have to estimate novel stimuli without feedback. Researchers in this field are interested in questions like how well (or if at all) participants pick up the environment, as abstraction-based theories assume, and how much people rely on their exemplar memory about what they have seen during the training (Juslin et al., 2008; Hoffmann et al., 2014; Pachur & Olsson, 2012). Researchers have interpreted the empirical results in favor of the view that people shifting between rule-based and exemplar-based strategies depending on several factors (Hoffmann et al., 2016; Juslin et al., 2008; Karlsson et al., 2007; Pachur & Olsson, 2012; Juslin, Olsson, & Olsson, 2003; Trippas & Pachur, 2019; Juslin, Jones, et al., 2003; Hoffmann et al., 2014).

The proposed cognitive models and empirical results presented in the first and the third paper directly challenge the idea that people use different strategies in different situations to form numeric estimations from multiple cues. In the first paper, we propose that presumed strategy shifts are the result of the additive combination of similarity-based memory retrieval and cue abstraction. In the third paper, we explicitly investigate whether people shift between or additively combine similarity-based and abstraction-based processes. The empirical results suggest that people indeed combine both processes additively.

Probability judgments and belief updating. In the field of probability judgments, researchers investigate how peoples' probability judgments deviate from normative solutions given by probability theory to better understand the cognitive processes underlying such judgments. One often-used paradigm in this domain is the so-called "book bag and poker chip" task (Edwards, 1968). In these tasks, participants receive possible hypotheses, for example, opaque book bags with different but explicitly stated proportions of differently colored poker chips. Given some evidence, for example a sample randomly drawn (with replacement) from one of the bags, participants are asked to state their belief that the sample stemmed from a specific bag. According to probability theory, subjective probability judgments, like the degree of belief that

a sample of chips stems from a specific bag, are calculated by estimating frequencies (e.g. Jenny, Rieskamp, & Nilsson, 2014), possibly with an associated error (Costello & Watts, 2014, 2018). More recently similarity-based theories for the formation of subjective probability judgments have been proposed (Read & Grushka-Cockayne, 2011; Nilsson et al., 2005).

The process of updating one's beliefs in the view of new information, for example a second sample drawn from the bag, is often modelled following a Bayesian approach (Tenenbaum, Griffiths, & Kemp, 2006; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Sanborn & Chater, 2016). An alternative, not-normative approach to belief updating are weighting-and-adding models. They propose that beliefs are combined by repeatedly updating currently held beliefs following an anchoring-and-adjustment process (Hogarth & Einhorn, 1992). Weighing-and-adding models, in contrast to Bayesian updating, can explain findings like the dilution effect, i.e. that consecutive probability judgments decrease in the face of non-diagnostic (non-informative) information, although they should remain the same (J. Shanteau, 1975; Nisbett, Zukier, & Lemley, 1981) and order effects, i.e. that the order of presentation has an impact on combined judgments (order effect; Hogarth & Einhorn, 1992; J. C. Shanteau, 1970). A finding related to the dilution effect which cannot be explained by weighing-and-adding models is the confirmation effect, i.e. that judgments increase in the face of non-diagnostic information if new evidence supports currently held beliefs (LaBella & Koehler, 2004).

The cognitive model proposed in the second paper explains how and why sequential judgments are increased (confirmed) or decreased (diluted) when non-diagnostic information is presented. The model assumes that subjective probabilities are the result of a similarity-based process and are updated in a weighing-and-adding process. The result of the weighing-and-adding processes is increased or decreased, depending on the similarity between a new piece of evidence and a held belief.

Integrating similarity-based memory retrieval and cue-abstraction

Albrecht, R., Hoffmann, J. A., Pleskac, T. J., Rieskamp, J., & von Helversen, B. (2020). Competitive retrieval strategy causes multimodal response distributions in multiple-cue judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(6), 1064-1090. http://dx.doi.org/10.1037/xlm0000772http://dx.doi.org/10.1037/xlm0000772

In the field of numeric estimation, researchers assume that people switch between similarity-based and abstraction-based processes depending on several extrinsic and intrinsic factors (Hoffmann et al., 2016; Juslin et al., 2008; Karlsson et al., 2007; Pachur & Olsson, 2012; Juslin, Olsson, & Olsson, 2003; Trippas & Pachur, 2019; Juslin, Jones, et al., 2003; Hoffmann et al., 2014). In this paper, we challenge this view by proposing a cognitive model that additively combines Cue abstraction with eXemplar memory (with exemplars activated based on similarities) assuming COMpetitive memory retrieval (hereafter CX-COM). We propose that the cognitive process underlying numeric estimation is a simple anchoring-and-adjustment process (Epley & Gilovich, 2001; Chapman & Johnson, 2002; Tversky & Kahneman, 1974), where the result of an exemplar retrieval is adjusted based on abstracted knowledge about the cue-criterion relationship. In contrast with the current view in exemplar theory, we assume that exemplar retrieval does not yield a weighted average of all stored exemplars' criterion values (Dougherty et al., 1999; Hintzman, 1984; Nosofsky, 1988) but instead propose that a single exemplar is recalled (Anderson, 1983; Logan, 1988; Nosofsky, 1997; Ratcliff, 1978). Specifically, given a to-be-judged object, in the first step *one* exemplar is retrieved from memory. The retrieval probability is calculated based on the same principles underlying established exemplar models - the more similar an exemplar is to the to-be-judged object, the more likely it is recalled. In a second step, the recalled criterion value is adjusted based on the difference in cue values between the to-be-judged object and the recalled exemplar. The smaller the difference between the cue-values, the smaller the adjustment. The adjustment is weighted cue-wise by importance learned during training. Clas-

sical exemplar models, in contrast, form a judgment by building an average of all exemplars in memory weighted by their similarity to the to-be-judged object.

We tested the assumptions underlying the CX-COM model theoretically and empirically. Analytically, we assessed the model's ability to predict the finding from the literature that people apply different strategies depending on the environment (Juslin et al., 2008; Hoffmann et al., 2016). Importantly, the mechanism proposed by CX-COM models this behavior without assuming strategy shifts. We reanalyzed the data presented by Juslin et al. (2008) to show that our model replicates response patterns from the literature, specifically that extrapolation behavior is mostly present in linear environments. Our model replicates these existing results well, suggesting that the supposed switch between different strategies automatically happens as a function of the environment and without the necessity for additional assumptions.

CX-COM assumes that similarity-based memory retrieval is competitive, meaning that always only one exemplar is recalled from memory. We assessed this assumption qualitatively. Under the usual assumption of a normally distributed error, competitive retrieval predicts that responses are multimodally distributed on a numeric scale. For example, consider participants have learned two exemplars during training, one with a high criterion value and the other with a low criterion value. If both exemplars are equally similar to the to-be-judged object they are both retrieved equally likely resulting in a bimodal response distribution. Even if one of the exemplars is more similar to the to-be-judged object, the other, less similar exemplar is still retrieved with a low probability resulting again in a bimodal response distribution with a higher peak at the response associated with the more similar exemplar.

In two experiments we tested the model quantitatively in a BIC-based (Bayesian Information Criterion; Schwarz et al., 1978) model comparison against three state-of-the-art competitors, classical exemplar models (Hoffmann et al., 2016), classical cue-abstraction models (Hoffmann et al., 2016), and a model that additively combines classical exemplar and cue-abstraction models (Bröder et al., 2017). In a training phase, participants had to learn six exemplars (five in the second experiment) by heart. An exemplar consisted of a visual presentation of three cues (four

in the second experiment) and an associated criterion value. During both phases, participants had to choose a value (from a scale from 1 to 33) for each presented exemplar and during the training, they received feedback stating the true criterion value of the exemplar. In the first experiment, we used a multiplicative environment and in the second experiment a linear environment.

In the second experiment, items were specifically chosen to test the qualitative prediction of multimodal response distributions. In general, even items with a very low retrieval chance should sometimes be retrieved. However, if the chance is very low for one exemplar to be retrieved it might not actually be retrieved in the limited number of trials per participant (15 in the first experiment and 10 in the second experiment). To show, in principle, that the response distributions are multimodal, we spread training items equally across the response scale. Four of the test items were very similar to two training items each; two with a large difference between the criterion values and two with a small difference. The model predicts clearly visible bimodal distributions over the response values for these four items. For the high-difference items, the modes of the bimodal distribution are predicted to be far apart and for the low-difference items they are predicted to be closer together.[1]

Overall, in the model comparison including four different models the CX-COM model was the most appropriate model for more than half of the participants and had the lowest mean BIC score. In the second experiment, participants showed clear signs of multimodal response distributions across and more importantly within subjects. In sum, CX-COM is a promising new model to explain the interaction between similarity-based and abstraction-based processes in a resource-efficient (Lieder, Griffiths, Huys, & Goodman, 2018b) and rational (Lieder, Griffiths, Huys, & Goodman, 2018a) way that allows researchers to derive distinct predictions for judgment behavior in various judgment situations.

[1]Note that the exact locations of the modes depend also on the cue-based adjustment mechanism.

Similarity-based weighing-and-adding for belief updating

Albrecht*, R., Jenny*, M.A., Nilsson, H., and Rieskamp, J. (2020b) The Similarity-Updating Model of Probability Judgment and Belief Revision. Manuscript under revision. *These authors contributed equally to this work.

In this paper, we propose a combination of similarity-based subjective probability judgments and a weighing-and-adding process for belief updating to explain several empirical findings from the literature on probability judgments. The proposed similarity-updating model is again based on an anchoring-and-adjustment mechanism (Epley & Gilovich, 2001, 2006; Chapman & Johnson, 2002; Tversky & Kahneman, 1974; Hogarth & Einhorn, 1992) and tested with a variant of the book-bag-and-poker-chip task. Participants received two decks of cards with three different colors and sequentially saw samples consisting of seven cards. The similarity-updating model assumes that single probabilities are the result of a process that compares how similar a sample is to a card deck. More precisely, the probability that a sample was drawn from a specific deck of cards is calculated by a softmax function and depends on the similarity between the sample and the deck relative to the similarities between the sample and the other deck (Luce, 1959). In the belief updating process when a second sample is presented, a weighted average of the two single probability judgments relative to one deck is calculated (Hogarth & Einhorn, 1992; Trueblood & Busemeyer, 2011) and used as an anchor. This average probability value is then adjusted by a similarity bias (von Helversen et al., 2014; Hahn et al., 2010) that increases or decreases the probability judgment as a function of the similarity difference between the first sample and a card deck, and the second sample and the card deck. More precisely, if the second sample is more similar to the considered deck than the first sample the probability judgment is increased, if it is less similar it is decreased (cf. LaBella & Koehler, 2004; Tentori, Crupi, & Russo, 2013).

We evaluated the model quantitatively and qualitatively in four experiments and against

a Bayesian model, and the Probability Theory plus Noise (PT+N) model (Costello & Watts, 2014, 2016, 2018). In the first experiment, participants received information about two decks of cards and each deck consisted of cards with three different colors. The information was given as the number of cards with a specific color out of a hundred. Self-paced, participants received one sample with seven cards and were asked to indicate, first, from which of the two decks the sample was drawn, and second, what the probability was that the sample stemmed from the indicated deck. Then they received a second sample (while all the previous information and choices remained visible on the screen) and were asked from which of the two decks *both* samples were drawn and what the associated probability was. Stimuli were selected specifically to test for the dilution effect and order effects. To this end, approximately 30% of trials included a sample that was equally likely drawn from both decks (i.e. a non-diagnostic sample). Each combination of samples was presented in both possible orders, meaning that each non-diagnostic sample was presented first in one trial and second in another trial. The other three experiments were meant as generalization tests for the findings from the first experiment. Stimuli were created based on the same principles and the experiments followed the same basic procedure. In the second experiment, the first sample was removed from the screen before the second sample was presented, thus, generalizing to situations where not all pieces of evidence are available at the same time. The information regarding the participant's choices, like the chosen deck and the associated probability, remained on the screen. In the third experiment people received a third sample, thereby generalizing to situations were beliefs have to be adjusted repeatedly. In the fourth experiment, evidence with two different sample sizes (7 cards or 14 cards) across trials was presented.

Qualitatively, we compared the model by considering the dilution effect, confirmation effects, order effects, converging evidence, and effects of sample size. The dilution effect, as well as its opposite the confirmation effect, namely that people increase their judgments in the light of non-diagnostic evidence, and order effects are predicted by the similarity-updating model and cannot be explained by normative models. The dilution effect is predicted by the nature of the

averaging mechanisms in the belief updating process (LaBella & Koehler, 2004). Confirmation effects are predicted by the similarity-based adjustment process and happen when the second sample is much more similar to the chosen deck than the first sample. Order effects are predicted by a recency parameter in the averaging mechanism (Hogarth & Einhorn, 1992). The Bayesian model and the PT+N model, on the other hand, predict that probability judgments become more extreme with increasing evidence in favor of one hypothesis (converging evidence) and increasing sample size. In all four experiments, we clearly find dilution effects in the majority of cases and confirmation effects in some cases. Also, a high proportion of order effects and only a smaller proportion of effects of converging evidence. In the fourth experiment, we find no effects on sample size. In the quantitative model evaluation, we compared the model to the Bayesian model and the PT+N model based on BIC. Overall, the similarity-updating model explains the data very well. It describes all but one or two participants best in each experiment and its median BICs are around one hundred points lower than for all other models.

All in all, the results speak in favor of the similarity-updating model. The model accounts both qualitatively and quantitatively very well for the data. Combining single probability judgments that are calculated through a similarity process with a weighing-and-adding process that is influenced by the similarity process composes a promising new theory for probability judgments and belief updating.

Coordinating similarity-based and abstraction-based strategies

Hoffmann, J A., Albrecht, R., von Helversen, B. (2020c) Coordinating several mental strategies requires integration: Evidence from human judgment. Manuscript submitted for publication.

In general, the idea that people switch between strategies has been widely accepted in the domain of judgment and decision making and factors leading to a switch between strategies have been examined in detail (Hoffmann et al., 2016; Juslin et al., 2008; Karlsson et al., 2007; Pachur

& Olsson, 2012; Juslin, Olsson, & Olsson, 2003; Trippas & Pachur, 2019; Juslin, Jones, et al., 2003; Hoffmann et al., 2014). However, the strategy coordination problem, namely how people combine or choose between different strategies to solve a problem, is still an unresolved question. In this paper we presented a new approach to investigate the strategy coordination problem by introducing a cognitive model called BASICS (**B**lending **A**nd **S**hifting **I**n **C**oordinating **S**trategies). On a theoretical level, we focused on three distinct forms of strategy coordination: strategy shifting, which assumes that people choose only one strategy to respond with, strategy blending, which assumes that they respond with a mixture of the different strategies' responses, and selective attention, which assumes that they respond with one strategy if there is a fitting one and respond with a mixture of not. As a method, we introduced a general version of mixture-of-experts learning models (Jacobs, Jordan, Nowlan, & Hinton, 1991; Jacobs, Jordan, & Barto, 1991). The idea behind mixture-of-experts models is that each strategy is represented separately by a so-called expert. Given a specific problem, each expert proposes a solution to that problem and a separate strategy-coordination mechanism integrates them (in case of blending) or chooses between them (in case of shifting). The experts learn how to apply their strategies in parallel with the strategy coordination mechanism that learns how to coordinate them.

We present an instance of BASICS for the domain of numeric estimation to investigate how similarity-based and abstraction-based mechanisms are combined. This model includes two experts, one implementing a rule-based strategy as a main effects regression model (Hammond & Stewart, 2001; Cooksey, 1996) and the other implementing a similarity-based exemplar model (Hoffmann et al., 2014; Juslin et al., 2008). The strategy selection mechanism is connected to the exemplars represented in the expert that implements exemplar memory and learns for which item to apply a rule and for which item to use an exemplar-based strategy. For example, if a to-be-judged object is very similar to an exemplar for which a rule-based strategy worked well in the past, then the strategy selection mechanism leans towards the rule-based strategy and vice versa. What it means that the strategy selection mechanism leans towards a strategy depends on two aspects, a general strategy preference and strategy specificity. The strategy preference

is a general bias towards one strategy. The strategy specificity directly impacts the type of strategy coordination. Assuming no general strategy preference, when the strategy specificity is very high, people shift between the different strategies even if there is only a small advantage for one strategy. If the strategy specificity is very low, they always build an average of both strategies responses, independently of the advantage one strategy might have. With a medium strategy specificity (selective attention), shifting only occurs if there is a clear advantage for one strategy.

We used BASICS to analyze how people coordinate strategies in the domain of numeric estimation. In a first step, we validated our model by training it in two linear and two non-linear environments. BASICS was able to predict the key findings from the literature that linear environments can be learned faster than non-linear environments and that extrapolation patterns are found in the linear but not in the non-linear environments. In a second step, we tested participants in two experiments using a so-called rule-plus-exception task (e.g. Kalish, Lewandowsky, & Kruschke, 2004). In the training phase, participants received items that followed a linear rule and items which were exceptions to that rule. In the training phase, we manipulated the frequency with which exceptions are encountered (10 times or 50 times) and in the test phase, we manipulated the similarity between test items and the learned rule and exception items. Across the two experiments, we manipulated if training items were repeated during the test (first experiment) or if only very similar items were presented during the test (second experiment).

During the test phase, we collected peoples' value and familiarity judgments (which of two exemplars seems more familiar). BASICS predicts that value judgments are more accurate for items that are similar to learned exception items if people use shifting (in both frequency conditions) or selective attention (in the high-frequency condition only). The reason is that with shifting the exemplar-based strategy is correctly used for exception items while with blending the response is a mixture of exemplar and rule-based strategies and therefore, the error is high even for items already encountered during training. For familiarity judgments, BASICS predicts that

the high-frequency exception items should be more familiar with selective attention compared to blending and shifting in order to selectively activate the correct strategy. The results in the two experiments provide a strong argument for strategy-blending compared to strategy-shifting and selective activation.

The BASICS model is a very general framework to model and analyze how different strategies and their coordination are learned. Instantiated in the domain of numeric estimation it allowed us to explore how similarity-based and abstraction-based strategies are coordinated. The results suggest that people additively combine different strategies and do not switch from one strategy to another between trials.

General Discussion

Similarity-based processes are an integral part of the judgment process and the question of how they are integrated sequentially and interact with human abstraction abilities has been hotly debated. In this dissertation, I contribute to this line of research in the domain of multiple-cue judgment with a cognitive modeling approach. On a theoretical level, the presented models provide alternative explanations for how cognitive processes interact. In the first and the second paper, I presented cognitive models that are based on simple heuristics, specifically anchoring and adjustment. These models show how judgments emerge from the simple processing of similarities and resource-efficient additive combinations of single information. In the third paper, I introduced an approach that defines a learning model used to investigate the very nature of strategy coordination by assuming that strategies are executed in parallel and asking how the results are combined to form a response.

The idea that people have a repertoire of strategies and combine them in some way or at least coordinate their use is broadly accepted, especially in the domain of judgment. However, there are still attempts to explain human cognitive processes in different domains using single-process theories. One major and very general example are Bayesian models of cognition, which are very

popular across several domains (for a review, see Chater, Oaksford, Hahn, & Heit, 2010). A Bayesian view on prototype models in categorization, for example, assumes that categories are represented by distributions (e.g. Gaussian) and the problem of categorizing an item comes down to finding the distribution that generated it (Chater et al., 2010). In contrast, until recently two groups of researchers debated if the cognitive process underlying categorizations can solely and exclusively be described with either exemplar processes (Nosofsky & Zaki, 2002; Nosofsky & Johansen, 2000) or based on prototypes (Smith & Minda, 2000). In reasoning, there is a long-standing debate on whether people exclusively use logical rules (Rips, 1994; Braine & O'Brien, 1998) or mental models (Johnson-Laird, 1989, 2006; Byrne, 2007) to solve a reasoning problem. In this dissertation, I take on the view that the solution to a problem, specifically the problem to form a numeric or probability judgment from multiple sources of information, is the result of the application and integration of several cognitive processes or systems (Ashby & Ell, 2002; Erickson & Kruschke, 1998; Evans, 2003; Hayes & Broadbent, 1988). Especially in the CX-COM model, the processes and their combination are deterministic and not post-hoc (Melnikoff & Bargh, 2018). The response is solely provided by the sketched anchoring-and-adjustment mechanism and different behaviors (formerly viewed as evidence for a shift between strategies) are the result of environmental properties alone.

Additive integration of information. Two cognitive models proposed in this dissertation, CX-COM and the similarity-updating model, use an anchoring-and-adjustment process to form a judgment. Anchoring and adjustment has mostly been investigated in the context of cognitive biases and describes the process of how information that is internally or externally accessible influences judgments (Tversky & Kahneman, 1974; Chapman & Johnson, 2002; Epley & Gilovich, 2001, 2006; Simmons, LeBoeuf, & Nelson, 2010). Although rare, there are models that explain the cognitive processes underlying judgment and decision making with anchoring and adjustment. For example, Hogarth and Einhorn (1992) proposed a belief-adjustment model where an initial belief is updated in the view of new evidence by a weighted adjustment relative to

some reference point. Millroth, Guath, and Juslin (2019) propose a risky choice model based on anchoring and adjustment as an alternative to Cumulative Prospect Theory. In their model, the anchor is likely given by the outcome that is more easily represented and this outcome is then adjusted as a function of the probability.

CX-COM and the similarity-updating model differ from other anchoring-and-adjustment models by assuming that both the anchor and the adjustment are the result of an internal cognitive process. In the CX-COM model, the anchor is the result of a memory retrieval process, where one exemplar is retrieved with a probability proportional to its similarity to the to-be-judged object (cf. Nosofsky, 1997). The adjustment is the result of a cue-abstraction process (cf. Hammond & Stewart, 2001) where the retrieved item's judgment value is changed in accordance with the difference between the to-be-judged object and the retrieved exemplar. In the similarity-updating model, single probabilities are calculated based on the similarities between a hypothesis and a piece of evidence. When probabilities are combined across different pieces of evidence, the anchor is the result of an additive combination of single probabilities (Hogarth & Einhorn, 1992) and this value is adjusted relative to the difference between the two pieces of evidence to the preferred hypothesis.

Anchoring and Adjustment and Mixtures of Experts. I presented two approaches in this dissertation to understand how similarity-based and abstraction-based processes (or strategies) interact in the domain of numeric estimation and they are different on a theoretical level. While the CX-COM model assumes an anchoring-and-adjustment (Chapman & Johnson, 2002; Epley & Gilovich, 2001) process, BASICS implements a mixture-of-experts model (Jacobs, Jordan, Nowlan, & Hinton, 1991; Jacobs, Jordan, & Barto, 1991). According to the mixture-of-experts approach, two (or several) strategies are executed fully, in parallel, and independently from each other. A response is some combination of the single strategies' responses. The anchoring-and-adjustment approach taken in CX-COM differs substantially as it assumes that strategies are calculated sequentially and that information from previous execution steps are used in the

application of later steps. More concretely, the cue-abstraction process which is applied second does not need to be calculated fully but only with respect to the results of the memory retrieval and the item to be judged.

In terms of resource management and cognitive cost the CX-COM model clearly outperforms BASICS. It has recently been shown that anchoring and adjustment is resource-rational and achieves a near-optimal speed-accuracy tradeoff (Lieder et al., 2018b, 2018a). Firstly, the anchor in the CX-COM model is the result of a memory retrieval where only one item is recalled from memory. While all items are activated in parallel, only one item is actually recalled and attended upon (Anderson & Crawford, 1980). In classical exemplar theory, in contrast, the judgment is the result of all criterion values stored in memory. This implies that all exemplars need to be attended to in some way either in parallel (see Dougherty et al., 1999; Hintzman, 1984) or consecutively (Nosofsky, 1997; Ratcliff, 1978). Secondly, the adjustment is only based on the weighted difference between the retrieved exemplar and the item at hand. Classical abstraction-based theories in the domain of judgment assume that the absolute importance of each cue and an intercept have correctly been inferred from the training items (Hoffmann et al., 2014; Juslin, Jones, et al., 2003). In contrast, the cue weights in CX-COM's cue-abstraction mechanism only reflect the relative importance of cues, similar to the attention weights applied in exemplar models (Nosofsky, 1986).

The BASICS model has two crucial advantages. Firstly, by modeling the learning process itself BASICS takes an additional process measure into account. Secondly, it is defined in very general terms and without the need for additional assumptions on the combination of processes. This implies that some of the disadvantages of resource management could be compensated. For instance, the exemplar-based strategy implemented in BASICS could easily be exchanged by an exemplar strategy assuming competitive retrieval, by only considering the criterion value of the exemplar with the highest activation. Furthermore, the general assumption that all strategies are fully executed could be weakened by penalizing the application of an increasing number of strategies during learning (Lieder, Callaway, Gul, Krueger, & Griffiths, 2017).

Although BASICS' defines a general architecture, the CX-COM model can only partly be implemented in BASICS. While the exemplar strategy can easily be adjusted to accommodate competitive retrieval, the cue-based adjustment process cannot be implemented straightforwardly. The adjustment is calculated relative to the retrieved exemplar, which implies that there would need to be a direct connection between the two strategies. However, this is at odds with one of the main assumptions of mixture-of-experts models, namely that different strategies calculate their responses independently of one another. On a behavioral level, CX-COM and BASICS as instantiated in the judgment domain predict similar error rates for items that are exceptions to a learned rule. In BASICS, blending predicts a high judgment error for exceptions because there is no correct switch to exemplar memory. The response is always a mixture of exemplar-based and rule-based processes. CX-COM also predicts a high error for exception items, because the exception items stored in memory would not always be recalled. One possibility to discriminate between the two models is to experimentally manipulate the distance between the rule and the exceptions during training. CX-COM predicts that the error rate decreases with increasing distance between rules and exemplars because the correct exception would more likely be recalled. Analytically, one could analyze the response distribution. CX-COM predicts that the error is sometimes low and sometimes high, depending on the retrieved item. BASICS with blending always predicts an average error.

Conclusion. In this dissertation, I proposed that similarity-based processes are an integral part of the judgment process and that they are sequentially integrated and interact with abstraction-based processes in a weighted, additive fashion. To this end, I presented three cognitive models that additively combine cognitive processes alongside with empirical evidence that favors these models' theoretical assumptions. Thus, this dissertation provides new and interesting insights into the judgment process and directly challenges the current view that people actively switch between different strategies.

References

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, *120*(1), 3–19.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*(3), 261–295.

Anderson, J. R., & Crawford, J. (1980). *Cognitive psychology and its implications*. Worth Publishers.

Ashby, F. G., & Ell, S. W. (2002). Single versus multiple systems of learning and memory. *Stevens' handbook of experimental psychology: Methodology in experimental psychology*, *9*(1), 655-691.

Attneave, F. (1950). Dimensions of similarity. *The American Journal of Psychology*, *63*(4), 516–556.

Braine, M., & O'Brien, D. P. (1998). *Mental logic*. Psychology Press.

Bröder, A., Gräf, M., & Kieslich, P. J. (2017). Measuring the relative contributions of rule-based and exemplar-based processes in judgment: Validation of a simple model. *Judgment and Decision Making*, *12*(5), 491–506.

Byrne, R. M. (2007). *The rational imagination: How people create alternatives to reality*. MIT Press.

Chapman, G. B., & Johnson, E. J. (2002). Incorporating the irrelevant: Anchors in judgments of belief and value. *Heuristics and biases: The psychology of intuitive judgment*, 120–138.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley*

Interdisciplinary Reviews: Cognitive Science, *1*(6), 811–823.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications.* Academic Press.

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, *121*(3), 463-480.

Costello, F., & Watts, P. (2016). People's conditional probability judgments follow probability theory (plus noise). *Cognitive Psychology*, *89*, 106–133.

Costello, F., & Watts, P. (2018). Probability theory plus noise: Descriptive estimation and inferential judgment. *Topics in Cognitive Science*, *10*(1), 192–208.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 968–986.

Dougherty, M. R., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review*, *106*, 180–209.

Edwards, W. (1968). Conservatism in human information processing. *Formal Representation of Human Judgment*, 17–52.

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science*, *12*(5), 391–396.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science*, *17*(4), 311–318.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107–140.

Evans, J. S. B. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459.

Evans, J. S. B. (2019). *Hypothetical thinking: Dual processes in reasoning and judgement.* Psychology Press.

Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, *143*(1), 182–194.

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT Press.

Glöckner, A., & Bröder, A. (2011). Processing of recognition information and additional cues: A model-based analysis of choice, confidence, and response time. *Judgment and Decision Making*, *6*(1), 23–42.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591–635.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.

Hahn, U. (2014). Similarity. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(3), 271–280.

Hahn, U., & Chater, N. (1998a). Similarity and rules: distinct? exhaustive? empirically distinguishable? *Cognition*, *65*(2-3), 197–230.

Hahn, U., & Chater, N. (1998b). Understanding similarity: a joint project for psychology, case-based reasoning, and law. *Artificial Intelligence Review*, *12*(5), 393–427.

Hahn, U., Prat-Sala, M., Pothos, E. M., & Brumby, D. P. (2010). Exemplar similarity and rule application. *Cognition*, *114*(1), 1–18.

Hammond, K. R., & Stewart, T. R. (2001). *The essential brunswik: Beginnings, explications, applications*. Oxford University Press.

Hampton, J. A. (2000). Concepts and prototypes. *Mind & Language*, *15*(2-3), 299–307.

Hayes, N. A., & Broadbent, D. E. (1988). Two modes of learning for interactive tasks. *Cognition*, *28*(3), 249–276.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*(2), 231–237.

Hintzman, D. L. (1984). Minerva 2: A simulation model of human memory. *Behavior Research*

Methods, Instruments, & Computers, *16*(2), 96–101.

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2014). Pillars of judgment: How memory abilities affect performance in rule-based and exemplar-based judgments. *Journal of Experimental Psychology: General*, *143*(6), 2242–2261.

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2016). Similar task features shape judgment and categorization processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(8), 1193–1217.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1–55.

Jacobs, R. A., Jordan, M. I., & Barto, A. G. (1991). Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks. *Cognitive Science*, *15*(2), 219–250.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive Mixtures of Local Experts. *Neural Computation*, *3*(1), 79–87.

Jenny, M. A., Rieskamp, J., & Nilsson, H. (2014). Inferring conjunctive probabilities from noisy samples: Evidence for the configural weighted average model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 203–217.

Johnson-Laird, P. N. (1989). *Mental models.* The MIT Press.

Johnson-Laird, P. N. (2006). *How we reason.* Oxford University Press, USA.

Juslin, P., Jones, S., Olsson, H., & Winman, A. (2003). Cue abstraction and exemplar memory in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 924–941.

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition*, *106*(1), 259–298.

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General*, *132*(1), 133–156.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of Linear Experts:

Knowledge Partitioning and Function Learning. *Psychological Review*, *111*(4), 1072–1099. doi: 10.1037/0033-295X.111.4.1072

Karlsson, L., Juslin, P., & Olsson, H. (2007). Adaptive changes between cue abstraction and exemplar memory in a multiple-cue judgment task with continuous cues. *Psychonomic Bulletin & Review*, *14*(6), 1140–1146.

LaBella, C., & Koehler, D. J. (2004). Dilution and confirmation of probability judgments based on nondiagnostic evidence. *Memory & Cognition*, *32*(7), 1076–1089.

Lewandowsky, S., Murdock, B. B., et al. (1989). Memory for serial order. *Psychological Review*, *96*(1), 25–57.

Lieder, F., Callaway, F., Gul, S., Krueger, P. M., & Griffiths, T. L. (2017). Learning to select computations. In *NIPS Workshop on Cognitively Informed AI* (Vol. 1711).

Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2018a). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, *25*(1), 322–349.

Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2018b). Empirical evidence for resource-rational anchoring and adjustment. *Psychonomic Bulletin & Review*, *25*(2), 775–784.

Little, D. R., & Lewandowsky, S. (2009). Beyond nonutilization: Irrelevant cues can gate learning in probabilistic categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 530–550.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*(4), 492–527.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. Wiley, NY.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*(2), 254–278.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238.

Melnikoff, D. E., & Bargh, J. A. (2018). The mythical number two. *Trends in Cognitive Sciences*,

22(4), 280–293.

Millroth, P., Guath, M., & Juslin, P. (2019). Memory and decision making: Effects of sequential presentation of probabilities and outcomes in risky prospects. *Journal of Experimental Psychology: General*, *148*(2), 304–324.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 775–799.

Nilsson, H., Olsson, H., & Juslin, P. (2005). The cognitive substrate of subjective probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(4), 600–620.

Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology*, *13*(2), 248–277.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 54–65.

Nosofsky, R. M. (1997). An exemplar-based random-walk model of speeded categorization and absolute judgment. *Choice, Decision, and Measurement: Essays in Honor of R. Duncan Luce*, 347–365.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, *7*(3), 375–402.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53-79.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(5), 924–940.

Pachur, T., & Olsson, H. (2012). Type of learning task impacts performance and strategy

selection in decision making. *Cognitive Psychology*, *65*(2), 207–240.

Raaijmakers, J. G., & Shiffrin, R. M. (1992). Models for recall and recognition. *Annual Review of Psychology*, *43*(1), 205–234.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Read, D., & Grushka-Cockayne, Y. (2011). The similarity heuristic. *Journal of Behavioral Decision Making*, *24*(1), 23–46.

Regehr, G., & Brooks, L. R. (1993). Perceptual manifestations of an analytic structure: The priority of holistic individuation. *Journal of Experimental Psychology: General*, *122*(1), 92–114.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. MIT Press.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Shanteau, J. (1975). Averaging versus multiplying combination rules of inference judgment. *Acta Psychologica*, *39*(1), 83–89.

Shanteau, J. C. (1970). An additive model for sequential decision making. *Journal of Experimental Psychology*, *85*(2), 181–191.

Simmons, J. P., LeBoeuf, R. A., & Nelson, L. D. (2010). The effect of accuracy motivation on anchoring and adjustment: Do people adjust from provided anchors? *Journal of Personality and Social Psychology*, *99*(6), 917–932.

Smith, D. J., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 3–27.

Söllner, A., Bröder, A., Glöckner, A., & Betsch, T. (2014). Single-process versus multiple-strategy models of decision making: Evidence from an information intrusion paradigm. *Acta Psychologica*, *146*, 84–96.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the

rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–665.

Sun, R. (1995). Robust reasoning: integrating rule-based and similarity-based reasoning. *Artificial Intelligence*, *75*(2), 241–295.

Sun, R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological Review*, *112*(1), 159–192.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318.

Tentori, K., Crupi, V., & Russo, S. (2013). On the determinants of the conjunction fallacy: Probability versus inductive confirmation. *Journal of Experimental Psychology: General*, *142*(1), 235–255.

Trippas, D., & Pachur, T. (2019). Nothing compares: Unraveling learning task effects in judgment and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(12), 2239–2266.

Trueblood, J. S., & Busemeyer, J. R. (2011). A quantum probability account of order effects in inference. *Cognitive Science*, *35*(8), 1518–1552.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

von Helversen, B., Herzog, S. M., & Rieskamp, J. (2014). Haunted by a doppelgänger. *Experimental psychology*, *61*(1), 12–22.

American Psychological Association

Journal of Experimental Psychology:
Learning, Memory, and Cognition

ISSN: 0278-7393

http://dx.doi.org/10.1037/xlm0000772

Competitive Retrieval Strategy Causes Multimodal Response Distributions in Multiple-Cue Judgments

Rebecca Albrecht
University of Basel

Janina A. Hoffmann
University of Bath and University of Konstanz

Timothy J. Pleskac
University of Kansas and Max Planck Institute for Human Development

Jörg Rieskamp
University of Basel

Bettina von Helversen
University of Zurich and University of Bremen

Research on quantitative judgments from multiple cues suggests that judgments are simultaneously influenced by previously abstracted knowledge about cue–criterion relations and memories of past instances (or exemplars). Yet extant judgment theories leave 2 questions unanswered: (a) How are past exemplars and abstracted cue knowledge combined to form a judgment? (b) Are all past exemplars retrieved from memory to form the judgment (integrative retrieval) or is the judgment based on one exemplar (competitive retrieval)? To address these questions we propose and test a new model, CX-COM (combining **C**ue abstraction with e**X**emplar memory assuming **COM**petitive memory retrieval). In a first step, CX-COM recalls only a single exemplar from memory. In a second step, the initially retrieved judgment is adjusted based on abstracted cue knowledge. Qualitatively, we show that CX-COM naturally captures judgment patterns that have been previously attributed to multiple strategies. Next, we tested CX-COM quantitatively in 2 experiments and found that it accounts well for people's judgment behavior. In the second experiment we additionally tested 2 qualitative predictions of CX-COM: The existence of multimodal response distributions within participants and systematic variability in judgments depending on the distance between similar exemplars in memory. The empirical results confirm CX-COM's assumptions. In sum, the evidence suggests that CX-COM is a viable new model for quantitative judgments and shows the importance of considering judgment variability in addition to average responses in judgment research.

Keywords: quantitative judgment, multiple cues, exemplar retrieval, cue abstraction, mixture models

Evaluating situations and judging the value of objects is a widespread cognitive task carried out every day in people's professional and private lives. From a judge passing sentence on a convict to a financial analyst evaluating the risk and value of a bond or stock, people's ability to estimate numerical criteria in many different domains is of high importance. When making judgments people use the information of different features or attributes (cues) describing an object or situation. A judge determining the length of a sentence for a robbery conviction, for example, might consider the extent of the damages in the case. To do so, the judge might retrieve details of past cases from memory and compare them with the facts of the current case. Such a

Rebecca Albrecht, Department of Psychology, University of Basel; Janina A. Hoffmann, Department of Psychology, University of Bath, and Department of Psychology, University of Konstanz; Timothy J. Pleskac, Department of Psychology, University of Kansas, and Center for Adaptive Rationality, Max Planck Institute for Human Development; Jörg Rieskamp, Department of Psychology, University of Basel; Bettina von Helversen, Department of Psychology, University of Zurich, and Department of Psychology, University of Bremen.

We thank Regina Weilbächer and Florian Seitz for their support in collecting the data and Anita Todd for proofreading the article. This research has been supported by the Swiss National Science Foundation (SNSF) Grant 146169 to Jörg Rieskamp and Bettina von Helversen and Grant 157432 to Bettina von Helversen. Rebecca Albrecht, Janina A. Hoffmann, Timothy J. Pleskac, Jörg Rieskamp, and Bettina von Helversen designed the research. Rebecca Albrecht, Janina A. Hoffmann, and Bettina von Helversen conceptualized the cognitive models and experimental design. Rebecca Albrecht performed the data analysis, implemented the experiments, and programmed the cognitive models. Rebecca Albrecht, Janina A. Hoffmann, and Bettina von Helversen wrote the original draft. Rebecca Albrecht, Janina A. Hoffmann, Bettina von Helversen, Timothy J. Pleskac, and Jörg Rieskamp edited and reviewed the article. The data and the models are available at: https://osf.io/96m8g/?view_only=278ea8c8fda34216b69f4a08328c7173.

Correspondence concerning this article should be addressed to Rebecca Albrecht, Department of Psychology, University of Basel, Missionsstrasse 62A, 4055 Basel, Switzerland. E-mail: rebecca.albrecht@unibas.ch

judgment strategy is usually described by exemplar models (Nosofsky, 2014). These models assume that people's judgments and decisions are based on the similarity between the object under consideration and exemplars stored in memory (Hoffmann, von Helversen, & Rieskamp, 2014; Juslin, Jones, Olsson, & Winman, 2003; Juslin, Olsson, & Olsson, 2003; Nosofsky, 1984, 1986, 1997). Exemplar models have been successfully used to explain a variety of phenomena across different domains ranging from memory recall (e.g., Brown, Neath, & Chater, 2007; Hintzman, 1984) to categorizations and classifications (Medin & Schaffer, 1978; Nosofsky, 1984) to decision making (Juslin & Persson, 2002; Pachur & Olsson, 2012; Platzer & Bröder, 2012). They have also been extended to account for judgments from multiple cues (Hoffmann et al., 2014; Hoffmann, von Helversen, & Rieskamp, 2016; Juslin et al., 2003; Juslin, Karlsson, & Olsson, 2008; von Helversen & Rieskamp, 2009).

Despite their success in cognitive psychology, approaches for quantitative judgments that are purely based on exemplar processing fail to address two problems: The first problem is that people learn to explicitly represent how cues relate to a criterion and use this knowledge to make predictions for new objects (Brehmer, 1994; Cooksey, 1996; Juslin et al., 2003). Following such a cue-abstraction process, the judge, returning to our earlier example, would pass a prison sentence in a robbery case by weighing the importance of the aggravating and mitigating factors (e.g., the damages caused and whether the robber showed remorse) and then combining the weighted factors to form a single sentence. Cue-abstraction processes are hard to reconcile with exemplar-based strategies. As a consequence, current research in judgment assumes that people rely on both processes but switch between them depending on the structure of the task and their own cognitive abilities (Herzog & von Helversen, 2018; Hoffmann et al., 2014, 2016; Juslin et al., 2003; Juslin et al., 2008; Juslin et al., 2003; Pachur & Olsson, 2012; von Helversen & Rieskamp, 2008, 2009). Yet, empirical evidence does not unequivocally favor the view that cue abstraction proceeds independently of exemplar retrieval. For instance, the similarity between the to-be-judged event and past instances influences people's judgments even when they are relying on rules and abstracted knowledge (Brooks & Hannah, 2006; Hahn, Prat-Sala, Pothos, & Brumby, 2010; von Helversen, Herzog, & Rieskamp, 2014). Still, it is an open question how people integrate the two types of processes, that is, cue-abstraction and exemplar-based processes, to form a judgment, which we address in the present work.

The second problem is that although exemplar models are deeply rooted in traditional models of memory, how exemplar models instantiate retrieval from memory diverges from the retrieval processes considered in contemporary memory models. Specifically, exemplar models in quantitative judgment assume an integrative retrieval mechanism where all previously encountered exemplars are activated in parallel and integrated into one composite value (Hoffmann et al., 2016; Pachur & Olsson, 2012). In contrast to this view, many contemporary memory models assume a competitive retrieval process where previously encountered exemplars compete for retrieval and only one exemplar is recalled (Anderson, 1983; Logan, 1988). The degree to which a competitive retrieval mechanism better captures how people retrieve past exemplars during the judgment process has not been investigated.

The goal of the present research was to propose and test an exemplar-based model for quantitative judgments that addresses these two limitations. CX-COM (combining **C**ue abstraction with e**X**emplar memory assuming **COM**petitive memory retrieval) proposes that people engage in a two-step judgment process: In the first step, people probabilistically retrieve one past exemplar from memory, and in the second step, they adjust the criterion value of the recalled exemplar based on knowledge about the cue–criterion relation. In the present work we first present evidence for different retrieval and knowledge integration mechanisms and their implications for the judgment process. Next, we formally derive the predictions of CX-COM from established versions of exemplar and cue-abstraction models and review how the new model can account for behavioral patterns frequently observed in multiple-cue judgment. We then present two experiments that (a) quantitatively test CX-COM against competing models (Experiment 1); and (b) test the qualitative prediction that previously learned exemplars compete for retrieval (Experiment 2). In the General Discussion section we compare CX-COM with cognitive models proposed for categorization and function learning and discuss limitations and potential future work.

Combining Exemplar and Cue-Abstraction Processes

Exemplar models propose that people represent learned instances by storing them in memory. Alternatively, cue-abstraction accounts propose that people conceptualize learned knowledge on a more abstract level as a set of rules or cue—criterion relations (Hahn & Chater, 1998; Macrae et al., 1998). The general question of how knowledge is represented and used has challenged judgment research over the past decade (Hoffmann et al., 2014, 2016; Juslin, Jones, et al., 2003; Juslin et al., 2008; Karlsson, Juslin, & Olsson, 2007; von Helversen, Herzog, & Rieskamp, 2014; von Helversen & Rieskamp, 2008, 2009). By analyzing judgment behavior in different domains, it has been shown that the judgment process varies with the structure of the judgment task and the cognitive abilities of the decision maker. For instance, participants are better described by cue-abstraction models if the judgment criterion is a linear, additive function of the cues, whereas nonlinear relationships are better captured with exemplar models (Hoffmann et al., 2016; Juslin et al., 2008). In this vein, current research in judgment portrays the judgment process as a selection from two types of judgment processes best described by rule-based cue-abstraction models or similarity-based exemplar models (Hoffmann et al., 2016; Juslin et al., 2008; Pachur & Olsson, 2012). This research implicitly suggests that people first select one process suited to the task at hand and then use only the output of this process to make judgments in the task.

Empirical evidence, however, suggests that exemplar retrieval and abstracted cue knowledge likely interact during categorization and judgment. Unintentionally activated exemplars can interfere with task performance if they do not match the demands of the current situation (e.g., Macrae et al., 1998). Specific exemplars can influence judgments that are otherwise based on abstracted cue knowledge (von Helversen et al., 2014) and activate different rules depending on the context in which past exemplars were learned (e.g., Yang & Lewandowsky, 2004). Consequently, categorization research tends to favor mixture or hybrid models that assume people's representations contain both generalized beliefs in the

form of abstracted (cue) knowledge and specific instances or exemplars. In these models both types of representations influence decisions, although the relative importance may differ depending on the task and learning history (e.g., Anderson & Betz, 2001; Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Erickson & Kruschke, 1998; Herzog & von Helversen, 2018; Nosofsky, Palmeri, & McKinley, 1994; Palmeri, Wong, & Gauthier, 2004; Vanpaemel & Storms, 2008).

Thus, it seems reasonable to assume that people integrate retrieval from memory with abstracted cue knowledge also in multiple-cue judgment. But how do these two processes interact? In categorization research different types of mixture and hybrid models have been proposed. Most prominently in blending models, an exemplar and a cue-abstraction mechanism process information in parallel and the two outputs are combined as a weighted average (e.g., Bröder, Gräf, & Kieslich, 2017; Erickson & Kruschke, 1998). The most recent implementation of a blending model for judgments is RulEx-J that captures the contribution of exemplar- and rule-processing across different task conditions (Bröder et al., 2017).

Integrative Versus Competitive Retrieval

Any model of memory has to address the key question of how exemplars stored in memory are retrieved, that is activated and recalled. Memory models often share the assumption that the activation of exemplars is based on their similarity to the current stimulus (the probe). They differ, however, in the way a recalled exemplar is produced upon request (Raaijmakers & Shiffrin, 1992). Two different approaches can be distinguished, an integrative and a competitive retrieval mechanism.

An integrative retrieval mechanism produces a composite of all exemplars (or of a subset) in memory. The most prominent example for an integrative retrieval mechanism is employed in the MINERVA model (Dougherty, Gettys, & Ogden, 1999; Hintzman, 1984). In this model, exemplars are represented as feature lists called memory traces. A probed recall activates all memory traces in parallel and yields a special memory trace: an echo. This echo is the sum of all traces in memory, each weighted by its activation value. Similarly, memory models that assume composite storage, such as TODAM (Lewandowsky et al., 1989), usually also yield a composite as a retrieval product.

If past exemplars compete for retrieval, only one exemplar is produced on each retrieval attempt. This competitive retrieval mechanism can be found in a number of established memory models, such as the ACT-R theory (Anderson, 1983), the instance theory of automatization (Logan, 1988, 2002), the search of associative memory (SAM) model (Raaijmakers & Shiffrin, 1980), and some random walk theories (Nosofsky, 1997; Ratcliff, 1978). A competitive retrieval mechanism assumes that exemplars in memory are stored and accessed separately. Although some of these theories specify how subsequently retrieved exemplars can be combined to form a task response, they assume that each retrieval request yields only one exemplar.

The retrieval mechanism employed (integrative vs. competitive) implies different response processes (Juslin & Persson, 2002; Palmeri, 1997). Exemplar models in categorization or judgment mostly postulate an integrative retrieval mechanism (e.g., Hoffmann et al., 2014; Juslin et al., 2003; Medin & Schaffer, 1978; Nosofsky, 1984; Pachur & Olsson, 2012). Once a probe is presented, all exemplars stored in memory are activated and a judgment or category response is formed as a weighted average over all memory items (Hoffmann et al., 2014; Juslin et al., 2003, 2008; Medin & Schaffer, 1978; Nosofsky, 1984). One potential reason why exemplar models seldom consider competitive retrieval is that often (at least in the domains that have been considered) the two mechanisms predict the same responses. In categorization tasks, for instance, classic exemplar models predict the probability of a new item belonging to a category by using the sum of similarities it holds with exemplars in that category. A competitive retrieval mechanism predicts that in each trial, an exemplar is recalled from memory and used as a basis for the category decision. However, the sum of the recall probabilities of individual exemplars belonging to a category is the same as the probability of assigning an item to a category if integrative retrieval is assumed. Thus, the two retrieval mechanisms cannot be distinguished in this type of task. In quantitative judgment tasks the type of retrieval can be important because integrative and competitive retrieval mechanisms make qualitatively distinct predictions on the judgment level.

Within the domain of judgments, integrative and competitive retrieval can be distinguished by the predicted trial-by-trial variability across items and (sometimes) different distribution shapes. An integrative retrieval mechanism predicts that all exemplars in memory are combined into a single response value. Within-participant trial-by-trial variability is typically assumed to be normally distributed (e.g., Pachur & Olsson, 2012; Pleskac, Dougherty, Rivadeneira, & Wallsten, 2009), resulting in a model-predicted *unimodal response distribution* centered around the predicted response value. With competitive retrieval each exemplar can, in principle, be recalled, although its chances in a given context might be very low. As a result, a competitive retrieval mechanism predicts *multimodal response distributions* and systematic changes in across-item variability. A detailed example will be discussed in the next section.

CX-COM: A Hybrid Model for Quantitative Judgment With a Competitive Retrieval Mechanism

The development of CX-COM was motivated by two currently unresolved questions in judgment research: (a) When forming a judgment based on exemplars retrieved from memory, does the retrieval request yield one exemplar or an integrative composite of all exemplars? (b) Does combining cue-abstraction processes with exemplar retrieval outperform the predictions of a pure exemplar-based or cue-abstraction process?

CX-COM addresses these two questions by proposing a two-step judgment process: In the first step, previously encountered exemplars compete for retrieval and only the winning exemplar along with its criterion value is recalled from memory. Second, a cue-based adjustment process uses the recalled exemplar as a reference point and adjusts the criterion value depending on the generalized beliefs about cue-criterion relations. Accordingly, CX-COM spells out how people may combine competitive exemplar-retrieval and cue-abstraction processes, allowing one to test these assumptions against single-process models as well as competing mixture models.

In this section, we first introduce established models of human judgment, that is, classical exemplar and cue-abstraction models.

Next, we explain CX-COM and its components in relation to the established models. We also introduce the blending model RulEx-J (Bröder et al., 2017) as an additional competitor. A running example at the end of each subsection highlights differences and similarities in the models' predictions. As a last step we review important findings from the judgment literature and explain how CX-COM accounts for them.

Exemplar Models

In exemplar models (Nosofsky, 2014), a probe p that has to be judged or categorized serves as a retrieval cue, activating previously encountered exemplars in memory. A response $\hat{j}_p^{\text{Exemplar}}$ is an average of all judgment values j_e associated with exemplars e in the set of all exemplars M weighted by their relative, subjective similarity to p,

$$\hat{j}_p^{\text{Exemplar}} = \frac{\sum_{e \in M} sim(e,p) \cdot j_e}{\sum_{e \in M} sim(e,p)}. \quad (1)$$

The more similar an exemplar in memory is to the probe, the higher its impact on the response value. The similarity between exemplars e and probe p depends exponentially on their distance in psychological space,

$$sim(e,p) = e^{-c \cdot dist(e,p)}. \quad (2)$$

Parameter c is the sensitivity parameter and manipulates how much impact the psychological distance between a probe and an exemplar has on the subjective perception of similarity. Lower values of c imply that two items with a high distance in psychological space are still perceived as similar.

The distance in psychological space is usually described using the family of Minkowski distance metrics. For an exemplar e with n cue dimensions and cue values $c_1^e, \ldots, c_n^e$, *a probe* p with cue values $c_1^p, \ldots, c_n^p$, and attention weights $w_1, \ldots, w_n$ this would be

$$dist(e,p) = \sqrt[r]{\sum_{i=1}^{n} w_i \cdot |c_i^e - c_i^p|^r}. \quad (3)$$

Attention weights are assumed to vary between 0 and 1 and are constrained to sum to 1. The parameter r captures how visually distinguishable cue dimensions are in a given task. The so-called city-block distance is defined by $r = 1$ and is used when the dimensions are very distinct (Garner, 2014; Shepard, 1964). The Euclidean distance is represented by $r = 2$ and is used when the dimensions overlap.

The response value $\hat{j}_p^{\text{Exemplar}}$ is a composite of the exemplars in memory and is predicted each time probe p is presented. Usually, a normally distributed error is associated with a response. Thus, the predicted distribution of criterion values, response distribution R_p^{Exemplar}, coincides with the assumed error distribution and is for one item p and some variance σ^2:

$$R_p^{\text{Exemplar}} \sim \mathcal{N}\left(\hat{j}_p^{\text{Exemplar}}, \sigma^2\right). \quad (4)$$

Competitive exemplar models make the same assumptions about the psychological distance (Equation 3) and similarity between the exemplars in memory and a probe (Equation 2). However, the relative similarity now determines the probability to recall exemplar e given probe p so that

$$pr(e \mid p) = \frac{sim(e,p)}{\sum_{e \in M} sim(e,p)}. \quad (5)$$

In each trial only one exemplar e is recalled from memory and the associated criterion value j_e is given as a response, that is, $\hat{j}_p^{\text{Exemplar-competitive}} = j_e$. In another trial, an exemplar with another criterion value may be retrieved probabilistically so that this competitive retrieval elicits a multimodal response distribution. Assuming also a normally distributed error associated with a response in every trial, the response distribution $R_p^{\text{Exemplar-competitive}}$ is a mixture of normal distributions with modes given by the criterion values j_e stored in memory:

$$R_p^{\text{Exemplar-competitive}} \sim \sum_{e \in M} pr(e \mid p) \cdot \mathcal{N}\left(\hat{j}_p^{\text{Exemplar-competitive}}, \sigma^2\right). \quad (6)$$

Thus, the predicted response distribution is quite different compared with the one predicted by the integrative exemplar model.

Example. To illustrate how an exemplar model with an integrative retrieval mechanism (Equation 4) and an exemplar model with a competitive retrieval mechanism (Equation 6) make different predictions in quantitative judgments, consider a judge passing sentence on a bank robber (see also Table 1). In an attempt to rob a bank, a robber (Defendant 1) caused low property damage and low harm to people. The judge can relate these circumstances to two earlier cases, one with low property damage (but high harm to people) and a prison sentence of 8 years, and another with low harm to people (but high property damage) and a sentence of 4 years. Assuming equal importance of both aspects of the case ($w_{\text{harm to people}} = w_{\text{property damage}}$), the sentence would be 6 years (because both old cases are equally similar to the new case). Assuming normally distributed deviations from the recalled criterion value in the model, repeated sentencing would result in a unimodal distribution of judgments centered around 6 years (see Figure 1, Exemplar (integrative)). In contrast, an exemplar model with a competitive retrieval mechanism (without additional assumptions on the response process) predicts sometimes a prison sentence of 4 years and sometimes a sentence of 8 years, depending on which of the two earlier cases the judge recalls. Assuming also that deviations from a recalled criterion value happen by chance, the judge would draw the prison sentence from a bimodal distribution with modes at criterion values 4 and 8 (see Figure 1, Exemplar (competitive)). Both integrative and competitive exemplar models would not predict a sentence below 4 years for Defendant 1 despite the fact that he caused less harm or less damage than the remembered convicts.

Consider in comparison Defendant 2 who caused high property damage and a high harm to people. Assuming equal dimension

Table 1
Bank Robber Example

	Memory		Novel	
Dimension	Convict 1	Convict 2	Defendant 1	Defendant 2
Property damage	High	Low	Low	High
Harm to people	Low	High	Low	High
Sentence	4	8	TBD	TBD

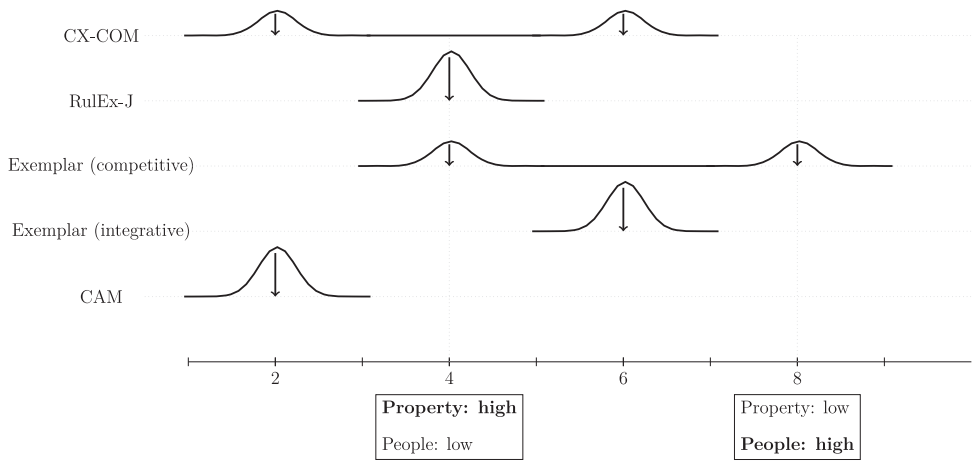


Figure 1. Example (see Table 1) showing the response distributions predicted by the cue-abstraction model (CAM), the exemplar model, RulEx-J, and the cue abstraction model with exemplar memory assuming competitive memory retrieval (CX-COM) for Defendant 1 (low property damage and low harm to people).

weights, both an integrative and a competitive exemplar model would predict the same sentence as for Defendant 1. The integrative exemplar models yields a unimodal sentence around 6 years; the competitive exemplar model a multimodal sentence, retrieving 4 and 8 years.

However, note that multimodal response distributions will not always occur with a competitive retrieval mechanism, but depend on the similarity structure of the training set given a probe, in particular the attention to specific dimensions. For example, assume again Defendant 2, who caused *high* harm to people and high property damage. This time the judge does not equally weight the two dimensions but only considers the harm to other people and neglects property damage ($w_{\text{harm to people}} = 100$, $w_{\text{property damage}} = 2$). In this case, Convict 2 will be perceived as much more similar to Defendant 2 than Convict 1 and will be retrieved with a much higher likelihood. As a result, the judge may always pass a sentence of 8 years and, consequently, a unimodal response distribution would emerge.

Cue-Abstraction Models

Cue abstraction models propose that people extract and explicitly represent their beliefs about the importance of cues as a set of weights. Cue-abstraction models (CAMs) for judgments are often implemented as main effects linear regression models (Juslin et al., 2008), because this implementation has been shown to fit judgment data especially well in a variety of domains (for reviews see Karelaia & Hogarth, 2008; Kaufmann, Reips, & Wittmann, 2013). To make a judgment, cue values $c_1 \ldots c_n$ of a probe p are weighted by their relative importance b_i and summed up so that

$$\hat{j}_p^{\text{CAM}} = k + \sum_{i=1}^{n} (b_i \cdot c_i). \quad (7)$$

Parameter k is an intercept, for example, the baseline judgment in case of cue values of zero. Similar to exemplar models, the CAM's prediction remains the same over repeated presentations of probe p and a normally distributed error is assumed resulting in the response distribution R_p^{CAM} centered around the response value $\hat{j}_p^{\text{CAM}}$ (similar to Equation 4).

Example. Figure 1 shows again an example of the unimodal response distribution predicted by the CAM in the bank robber's case (Table 1, Defendant 1). Assuming a minimum sentence of 2 years for a robbery attempt with low property damage and low harm to people, the judge could weigh high property damage as an additional 2 years of prison and high harm to people as an additional 6 years.[1] This would result in a prison sentence for the novel Defendant 1 of 2 years. For Defendant 2 (high property damage and high harm to people) the CAM with the same assumptions would thus predict a unimodal response distribution with a mode at 10 years.

Combining Competitive Exemplar Retrieval With Cue-Abstraction

With CX-COM we propose a two-step process: First, an exemplar is recalled from memory. Second, the associated criterion value is adjusted based on the beliefs about the cue–criterion relationship. Exemplar retrieval follows the same principles as in the exemplar model with a competitive retrieval mechanism. Following the presentation of a probe p, all exemplars e in exemplar memory M are activated based on their relative, subjective simi-

[1] Note that we use different dimension weights for the exemplar model and the CAM to illustrate how the mechanisms assumed by the models can lead to differential predictions. If we assume the same dimension weights in the exemplar model as in the CAM, i.e. $w_{\text{harm to people}} = 6$, $w_{\text{property damage}} = 2$, the exemplar models become more difficult to distinguish. The integrative exemplar model predicts a unimodal distribution centered around a prison sentence of 4.1 years for Defendant 1 and centered around 7.9 years for Defendant 2. The exemplar model with competitive retrieval predicts a bimodal distribution with modes at the two sentences of the previous cases four and eight. However, for Defendant 1 the judge has a 98% chance to recall the case of Convict 1 and for Defendant 2 he has a 98% to recall the case of Convict 2. For CX-COM we discuss how its predictions depend on its parameters in the section "Predicting multiple-cue judgments with CX-COM".

larity with similarity and psychological distance calculated as described in Equations 2 and 3. The relative similarity determines the probability to recall exemplar e and is described in Equation 1.

In each trial one exemplar is recalled from memory. This exemplar is used as a reference point for a cue-abstraction process. Specifically, a cue-based adjustment mechanism adjusts the criterion value j_e of a recalled exemplar e to obtain a response. The magnitude of the change depends on the differences in cue values between the probe p and the exemplar e on each cue dimension and the relative importance given to the cue dimension, represented by a cue dimension weight b_i:

$$\hat{j}_p^{\text{CX-COM}} = j_e + \left(\sum_{i=1}^{n} b_i \cdot \left(c_i^e - c_i^p\right)\right) \cdot \alpha. \tag{8}$$

Parameter α is a scaling parameter that reflects how much the observed difference in cue values influences the judgment, and c_i^p and c_i^e denote the cue values of the probe and the recalled exemplar in cue dimension i.

The competitive retrieval in CX-COM elicits a response distribution quite different from the integrative exemplar model and the cue abstraction model. Assuming a normally distributed error associated with a response in every trial, the response distribution $R_p^{\text{CX-COM}}$ is a mixture of normal distributions with modes close to criterion values j_e stored in memory and weighted by the similarity of the associated exemplars e to the probe p:

$$R_p^{\text{CX-COM}} \sim \sum_{e \in M} pr(e \mid p) \cdot \mathcal{N}\left(\hat{j}_p^{\text{CX-COM}}, \sigma^2\right). \tag{9}$$

Example. Figure 1 also shows an example of the multimodal response distribution predicted by CX-COM in the bank robber's case (Table 1, Defendant 1). Assuming equal importance of both aspects of the case, the similarity between Defendants 1's case and the two older cases is the same and so is their chance to be recalled, similar to the predictions of the exemplar model with competitive retrieval. In CX-COM, however, the modes of the multinomial response distribution are shifted depending upon the difference in cue values between the current case and the recalled case. Making the simplified assumption that a difference between high and low damage/harm is 2 years of prison, the modes would be adjusted downward by 2 years each from 4 years to 2 years and from 8 years to 6 years, respectively. For Defendant 2, however, the predicted modes are adjusted upward from 4 to 6 years and from 8 to 10 years, respectively. Thus, because the cue abstraction component is sensitive to the direction of the adjustment, CX-COM also predicts different response distributions from an exemplar model with only competitive retrieval.

Blending Models

Besides, pure exemplar and cue abstraction models, blending models such as ATRIUM for categorization (Erickson & Kruschke, 1998) and the measurement model RulEx-J for judgments (Bröder et al., 2017) have been proposed. These models assume an independent processing of exemplar and cue-abstraction models with the overall response being a weighted average (or blend) of the predictions of the single responses. As it is the most recent blending model for judgments, we included RulEx-J in the model test.

Given the judgments for probe p predicted by the exemplar model, $\hat{j}_p^{\text{Exemplar}}$, and the CAM, $\hat{j}_p^{\text{CAM}}$, the response of RulEx-J is

$$\hat{j}_p^{\text{RulEx-J}} = \beta \cdot \hat{j}_p^{\text{CAM}} + (1 - \beta) \cdot \hat{j}_p^{\text{Exemplar}} \tag{10}$$

with the parameter β weighting the relative contribution of each model's response. The response distribution $R_p^{\text{RulEx-J}}$ is unimodal as in the exemplar model and the CAM, but the mode lies between the predictions of these models:

$$R_p^{\text{RulEx-J}} \sim \mathcal{N}\left(\hat{j}_p^{\text{RulEx-J}}, \sigma^2\right). \tag{11}$$

Although they are both mixture models, the blending model differs from CX-COM in two important aspects: (a) the blending model assumes a parallel processing of an exemplar and a cue-abstraction component while CX-COM assumes that cue abstraction acts upon the retrieved exemplar, and (b) the blending model's exemplar component assumes integrative retrieval while CX-COM assumes competitive retrieval.

Example. According to RulEx-J, the judge sentences Defendant 1 (see Table 1) to a mixture of the predictions of the Exemplar model with integrative retrieval and the CAM. Making the same assumptions for the two models as in the respective examples and additionally assuming that the judge gives equal weights to both models' predictions, the sentence would be 4 years, exactly the middle between the exemplar model's prediction (6 years) and the CAM's prediction (2 years).

Predicting Multiple-Cue Judgments With CX-COM

In the past decade, research on multiple cue judgments has proposed that judgment strategies may elicit distinct behavioral judgment patterns (Bröder et al., 2017; Hoffmann et al., 2014, 2016; Juslin et al., 2008; Mata, von Helversen, Karlsson, & Cüpper, 2012; Pachur & Olsson, 2012; von Helversen, Mata, & Olsson, 2010; von Helversen & Rieskamp, 2009). For instance, Juslin, Karlsson, and Olsson (2008) showed that in a linear judgment task people showed extrapolation, that is they judged probes with lower/higher cue values than the training exemplars as having lower/higher criterion values than the training exemplars. This judgment pattern matches the predictions of the CAM (Figure 2a), but disagrees with the predictions of an exemplar model in that task (Figure 2b). In contrast, in a multiplicative environment participants do not seem to extrapolate beyond the range of encountered training values and thus participants' responses match the predictions of the exemplar model (Figure 2e), but disagree with the CAM's predictions (Figure 2d). In general, these differences in judgment patterns have been taken as evidence that people shift between exemplar memory and cue abstraction processes (Bröder et al., 2017; Hoffmann et al., 2014, 2016; Juslin et al., 2008; von Helversen et al., 2010).

CX-COM does not assume a shift between judgment processes, but proposes that a cue adjustment process acts on the retrieved exemplars. The α parameter governs the extent to which the retrieved criterion value is adjusted based on cue knowledge. In the following we show that CX-COM can capture the same behavioral judgment patterns that have been reported in the literature without assuming a change in judgment processes and analyze how different parameter settings influence CX-COM's predictions.

In a first step, we generated CX-COM's predictions for the linear and the multiplicative judgment task reported by Juslin et al. (2008), using the reported parameters for the CAM as dimension weights and an additive similarity function with equivalent param-

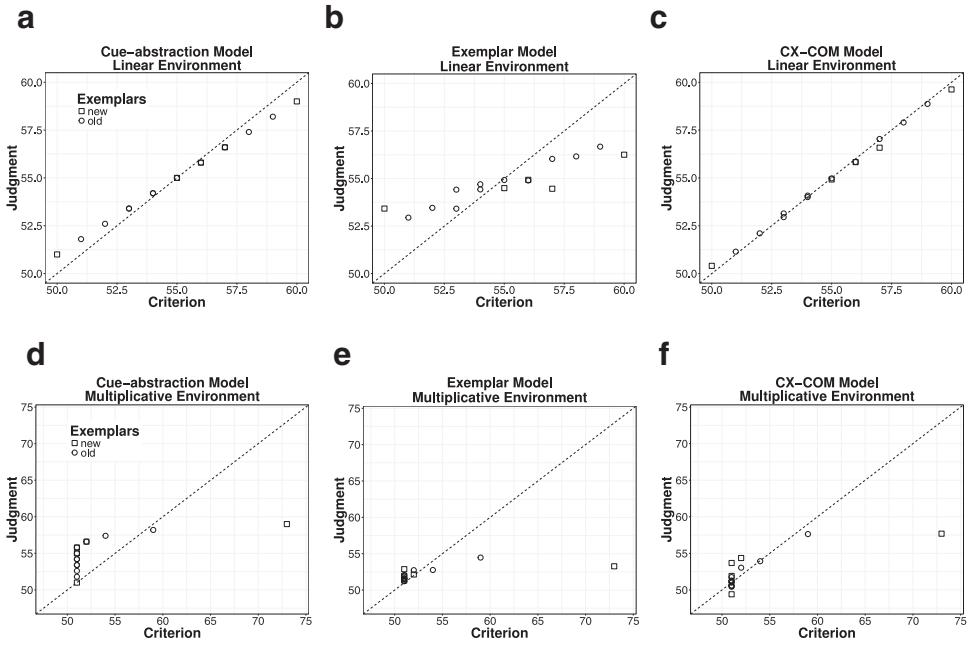


Figure 2. Shows predictions of a cue-abstraction model (panels a and d), an exemplar model (panels b and e) and CX-COM (panels c and f) in a linear environment (panels a–c) and a multiplicative environment (panels d–f). Items, models, and parameters are the same as described in Juslin et al. (2008). For CX-COM we used the same weights as for the linear model ($w_1 = 3.2$, $w_2 = 2.4$, $w_3 = 1.6$, $w_4 = 0.8$) and $\alpha = 1$.

eters as attention weights for the exemplar model. Figure 2c and 2f illustrate that CX-COM predicts a similar change in judgments depending on the task structure as predicted by a strategy shift. In the linear environment, its predictions resemble the predictions of the CAM, whereas its predictions lie between the exemplar model and the CAM in the multiplicative environment. Thus CX-COM reflects participants' responses in both environments. Notably, CX-COM accounts for these behavioral patterns without adjusting any parameter values across environments but the changes result from differences in the structure of the environment. One reason is that in a linear task with correct weights the adjustment process by CX-COM leads to the same judgment independent of which exemplar was retrieved.[2] In contrast, in a multiplicative task predictions will differ depending on the retrieved exemplar leading to exemplar effects and reducing extrapolation on the average level.

But can CX-COM also explain strategy shifts within the same environment due to within-task manipulations such as instructions or individual preferences? To understand whether the free parameters in CX-COM allow it to capture exemplar-based and cue-based judgment patterns within a task, we analyzed within Juslin et al.'s (2008) linear environment (Juslin et al., 2008) how changes in the parameter values, specifically changes in the adjustment parameter α and in the dimension weights, influence CX-COM's predictions. Assuming correct dimension weights with parameter $\alpha = 1$ (Figure 3e), CX-COM produces the exact same predictions as the CAM (Figure 3a), and CX-COM with $\alpha = 0$ (Figure 3f) produces the exact same predictions as the exemplar model (Figure 3b). Assuming incorrect weights, for instance uniform weights equal to 1, the predictions of CX-COM for $\alpha = 0$ (Figure 3h) are still exactly the same as for the exemplar model (Figure 3d). However, with $\alpha = 1$ the predictions differ between the CAM (Figure 3c) and CX-COM (Figure 3g), with CX-COM showing a judgment pattern that deviates from the CAM with incorrect weights, but is still linear. Accordingly, on average, if $\alpha = 0$ CX-COM reduces to an exemplar model with the same dimension weights. With increasing α,[3] that is, more cue adjustment, the predictions become more linear resembling a rule-based process. This suggests that α in CX-COM reflects the extent to which participants' judgments are influenced by cue-abstraction processes, similar to the interpretation of β in other mixture models like RulEx-J. However, the CAMs and CX-COMs predictions are only identical when the correct cue weights are assumed.

These simulations indicate that CX-COM can also reflect different levels of exemplar and cue-abstraction processes induced by manipulations within a task environment. For instance, previous research has argued that changing only one cue between subsequent exemplars facilitates cue abstraction processes (Juslin et al., 2008). Within CX-COM, this could be reflected by a stronger reliance on adjustment processes resulting in a lower α parameter for confounded than for ordered sequences.

Overall, the simulations demonstrate that CX-COM is able to account for important empirical findings in the judgment literature with the α parameter reflecting different levels of cue-abstraction

[2] For example, assume training items ($c_1 = 1$, $c_2 = 1$) with Criterion 4 and ($c_1 = 2$, $c_2 = 2$) with Criterion 8. Assume further that a participant abstracts the (correct) cue weights $w_1 = w_2 = 2$ from these items. For a test item ($c_1 = 1$, $c_2 = 2$) the response would be 6, independently of whether the first training item is recalled and the associated criterion value is increased or if the second training item is recalled and its criterion value is decreased.

[3] Note that α as well as the dimension weights depend on the judgment scale and thus cannot be easily compared between tasks.

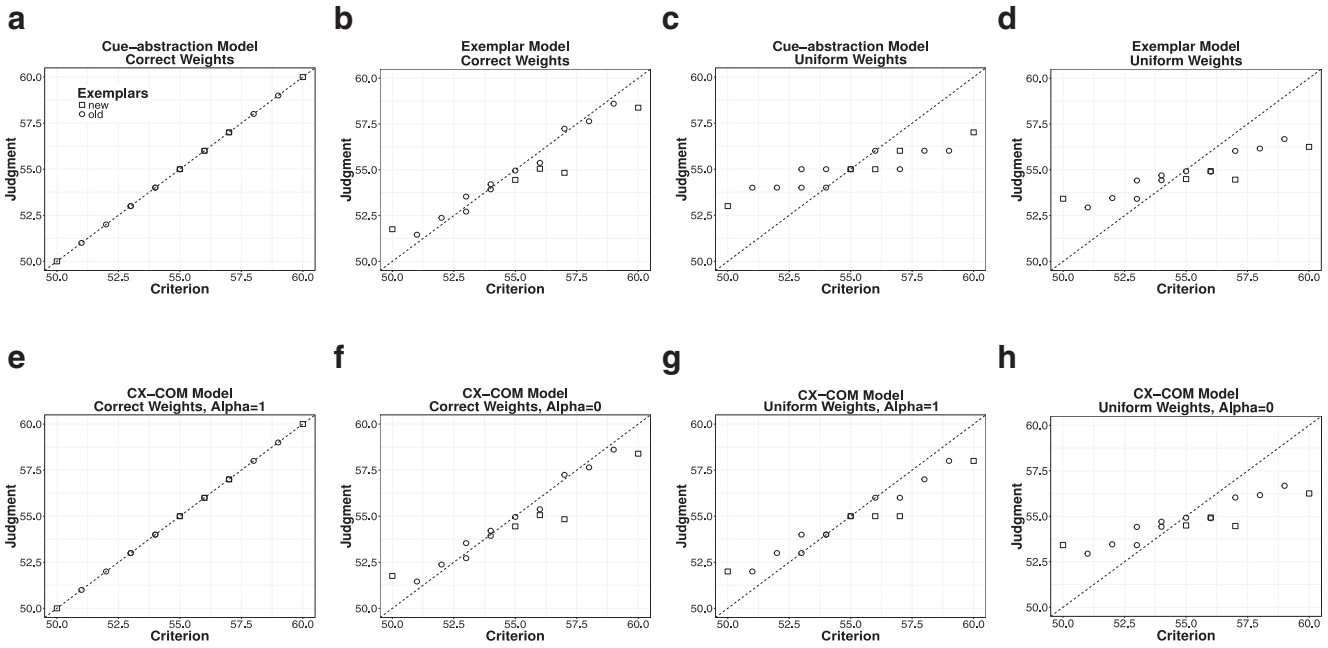


Figure 3. Predictions of the CAM (panels a, c), the exemplar model (panels b, d) and CX-COM (panels e–h) assuming correct weights ($w_1 = 1$, $w_2 = 2$, $w_3 = 3$, $w_4 = 4$; panels a, b, e, d) and incorrect uniform weights ($w_1 = 1$, $w_2 = 1$, $w_3 = 1$, $w_4 = 1$; panels c, d, g, h). Sensitivity Parameter c in all models including an exemplar component is set to the sum of the weights and attention weights are set to the sum of weights divided by c. Panels e and g show CX-COM's predictions assuming a strong influence of the cue-abstraction process (i.e., $\alpha = 1$) and panels f and h shows CX-COM's predictions assuming a pure exemplar process (i.e., $\alpha = 0$). When assuming correct weights CX-COM perfectly mimics the exemplar model's and the CAM's predictions depending on the value of parameter α (compare panels a and e, and panels b and f). Assuming uniform weights CX-COM with a α of 0 still matches the exemplar model's predictions (compare panels d and h), however, the predictions differ between the CAM and the CX-COM model with $\alpha = 1$ (compare panels c and g).

processes.[4] Although these results suggests that CX-COM is more flexible than either the CAM or the exemplar model, it still provides a more parsimonious explanation than a blending model like RulEx-J (Bröder et al., 2017).

Testing CX-COM's New Predictions of Judgment Behavior

In the previous section we showed that CX-COMs can capture patterns of judgments reported in the literature and usually attributed to a shift in judgment strategies. However, the data of these studies does not allow comparing CX-COM with the other models because usually each item is only repeated once or twice making it impossible to distinguish the models. The reason is that CX-COM's unique characteristic is the shape of the response distribution and thus it only makes different predictions if an item is repeated many times. Accordingly, we conducted two new experiments to quantitatively and qualitatively test CX-COM's predictions.

Quantitative test. CX-COM combines the judgment processes of two very well established cognitive models, the CAM and the exemplar model. To quantitatively test CX-COM, we compared it against several competitors: an exemplar model, a CAM, RulEx-J, and a baseline model (for details on a recovery study with previous data see Appendix A). The baseline model provides a benchmark for the absolute fit of the model. In the baseline model, we assume that participants respond with a constant value (with added noise), that is, we fit a normal distribution with mean and variance as free parameters to participants. To better take CX-COM's functional flexibility into account, we also did a cross-validation for both experiments. All details concerning the fitting procedure and mean parameter values are shown in Appendix B.

Qualitative test. The CX-COM model predicts judgment patterns that are qualitatively distinct from single exemplar and cue-abstraction models. Specifically, CX-COM's competitive retrieval mechanism predicts multimodal response distributions and systematic changes in variability across items. This is in stark contrast to the classical exemplar models with integrative retrieval and the CAM which always predict a unimodal distribution centered around one model-predicted value. We tested the assumption of a competitive retrieval process explicitly in Experiment 2: A competitive retrieval process predicts that variations in judgments across and within items depend on the number of similar exem-

[4] Please note that a quantitative analysis of existing data in a more traditional paradigm does not allow to distinguish CX-COM from previously proposed judgment models. Due to a low number of observations per item the models are not recoverable. See Appendix A for more details.

plars in memory and the distance between criterion values for similar exemplars. If a probe activates only one similar exemplar, the variability should be lower than if several similar exemplars are activated. If several exemplars with similar judgment values are activated, the variability in judgments is low. But if a probe activates exemplars with strongly dissimilar criterion values, high judgment variability and multimodal response distributions are predicted.

Experiment 1

Experiment 1 was designed as a first, quantitative test for the CX-COM model. Specifically, we aimed at testing CX-COM in a situation that would usually favor an (integrative) exemplar model. In the experiment, participants had to solve a quantitative judgment task using three cues. To encourage exemplar retrieval, participants learned to judge a small set of training items and their criterion values by heart (Rouder & Ratcliff, 2006). After this training phase, they were instructed to repeatedly judge the criterion values of novel test items on the basis of their similarities to the training items.

Method

Participants. We tested 29 current or former students from the University of Basel (M_{age} = 27 years, SD = 7, range: 20–46 years). The target sample size was a priori set to 30 following conventions for one condition in cognitive modeling research (e.g., Hoffmann et al., 2016; Tsetsos et al., 2016). Thirty-five participants were invited through the recruitment platform of the center for Economic Psychology in Basel and 29 came at the assigned time. The experiment took on average approximately 1 hr. Participants could choose between course credit or a payment of 20 Swiss francs per hour. In addition, participants could earn a performance-dependent bonus of 5 Swiss francs. The study received ethics approval by the Institutional Review Board (IRB) of the Faculty of Psychology at the University of Basel.

Materials. In Experiment 1 we used items with three dimensions shown as three adjoining stacks (left, middle, right) on the left side of a computer screen, see Figure 4. Each dimension was assigned a value between 1 and 4, indicated by the number of geometric shapes in the stack. Additionally, the dimensions differed in geometric shape (triangle, square, circle) and color (blue, red, green). Colors were chosen as complementary colors from the color wheel rendering them all similarly visually salient. Associated criterion values ranged between 1 and 33 and were presented on a half-circle on the right side of the computer screen. Figure 4 shows the visual presentation of stimuli as shown to the participants. Positions of the different cue dimensions were randomized across participants. Throughout the text, items are named with their three cue values separated by a dot. Item 3.1.1, for example, corresponds to an item with value 3 in Cue Dimension 1 (i.e., three shapes in the left position) and 1 in Cue Dimensions 2 and 3 (i.e., one shape in the middle and the right position).

To foster the use of exemplar-based processes we used a multiplicative environment (Hoffmann et al., 2016; Juslin et al., 2008) and instructed participants to use similarities to judge novel items (Olsson, Enkvist, & Juslin, 2006). To help ensure that participants did not rely only on simple visual features of the items, for example, the higher the cue value the higher the criterion value, the first cue was inverted so that

$$j = (5 - c_1) \cdot c_2 \cdot c_3, \tag{12}$$

with cue values $c_1, \ldots, c_3$.

We presented only a small number of training items (six) that had to be learned by heart but used a larger number of novel test items (14) to evaluate participants' responses. The six training items were chosen such that the associated criterion values represented the general trend found in multiplicative environments; the lower part of the response scale was densely packed with observations (criterion values 4, 6, 8); in the higher part of the scale, single observations were rather sparse (criterion values 12, 18, 24).

Ten out of 14 test items were chosen so that each was most similar to one of the training items. The similarities are calculated

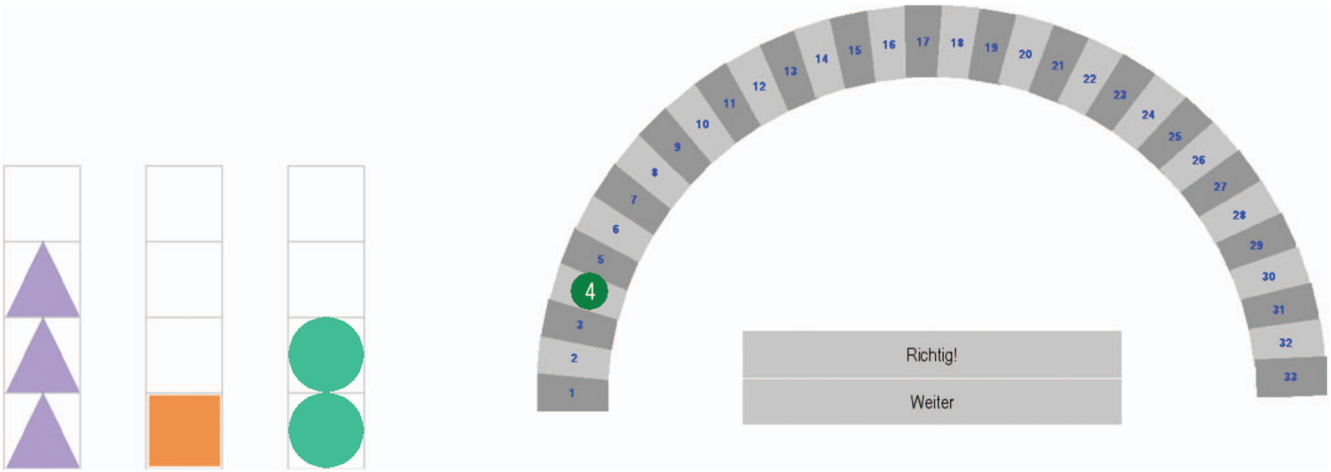


Figure 4. Visual presentation of stimuli in Experiment 1. The depicted stimuli 3.1.2 (on the left) is associated with criterion value 4 (highlighted on the half circle on the right). The text on the figure is German stating "Richtig!" for "Correct!" and "Weiter" for "Next." See the online article for the color version of this figure.

with the city-block distance metric and assuming equal dimension weights. Four additional test items were included as fillers for which we did not systematically vary/control the similarity to all training items (see Table 2). We chose a relatively small number of training items and larger number of test items for two reasons: First, we wanted direct control over the similarity structure among training items and between training and test items. Second, we wanted participants to learn the training items by heart and to remember them throughout the whole experiment.

Procedure. The experiment included three phases. In the training phase participants had to learn six different training items by heart. They were told that there was no simple functional dependency between the cues and criterion values and that they would later be asked to use the learned items to estimate the criterion values for novel items. We used two types of training blocks, judgment learning and cue learning. In the judgment-learning blocks participants were presented with the cue values of a training item on the left-hand side of the screen and had to choose the associated criterion value from the response circle on the right-hand side of the screen. In the cue-learning blocks one specific criterion value was highlighted on the response circle and participants were asked to adjust the stacks so that they represented the cues of the training item corresponding to the displayed criterion value. Participants could adjust the stack by repeatedly clicking on it to change the number of displayed shapes. After giving a response, participants received feedback in all trials during training. In total, the training phase consisted of 10 blocks.[5] Within each block the six training items were presented once in a random order. The training phase included six judgment-learning and four cue-learning blocks presented in alternating order, and with two judgment-learning blocks at the beginning and one judgment-learning block at the end of training. Participants received a bonus of 5 Swiss francs if they were able to correctly judge all training items in at least two subsequent training blocks.

Table 2
Stimuli, Manipulation, and Results in Experiment 1

	Training item						
	4	6	8	12	18	24	Result
Criterion test item	3.1.2	2.1.2	3.2.2	1.1.3	2.2.3	1.2.3	Mean (*SD*)
3.1.1	**1**	2	2	4	4	5	6.37 (4.13)
4.1.2	**1**	2	2	4	4	5	8.53 (5.59)
2.1.1	2	**1**	3	3	3	4	4.50 (2.79)
3.2.1	2	3	**1**	5	3	4	10.93 (6.66)
4.2.2	2	3	**1**	5	3	4	10.35 (4.68)
1.1.4	4	3	5	**1**	3	2	18.49 (4.82)
2.2.4	4	3	3	3	**1**	2	21.92 (3.71)
2.3.3	4	3	3	3	**1**	2	19.78 (6.47)
1.2.4	5	4	4	2	2	**1**	24.74 (2.96)
1.3.3	5	4	4	2	2	**1**	21.27 (5.86)
2.3.2	3	2	2	4	2	3	15.00 (5.98)
2.1.3	2	1	3	1	1	2	12.61 (4.47)
1.3.2	4	3	3	3	3	2	15.57 (5.54)
2.3.1	4	3	3	5	3	4	14.77 (5.62)

Note. Distance profiles for stimuli used in Experiment 1. Items consist of three cue dimensions (cf. Figure 4). Dimension values are separated by a dot. The distances are calculated using the city-block metric assuming attention weights of 1 in each dimension. Items with a distance of 1 to a test item are assumed to be most similar to that item and are shown in bold.

In the test phase, participants saw 14 different novel test items and were asked to estimate the associated criterion values. People were asked to make the judgment according to the test items' similarities to the training items. The test phase included 15 test blocks. In each block all 14 items were presented in random order, resulting in 210 test trials. After Test Blocks 5 and 10, participants again judged the criterion values for all training items twice, resulting in four additional judgment-learning blocks during the test phase. These blocks were announced as training blocks and participants received feedback.

After completing the test phase, participants judged on a paper-and-pencil questionnaire how similar each test item was to each of the six training items, resulting in a total of 84 similarity judgments. Each page of the questionnaire showed one test item and the six training items. Participants were asked to judge the similarity on a scale of 0 (*completely different*) to 10 (*exactly the same*) for each pair.[6] The results of the similarity questionnaire are presented in Appendix C.

Results

Performance. In the last two training blocks (one cue-learning block and one judgment-learning block) participants judged 80% of the training items correctly. In the four judgment training blocks in the test phase they judged 81% of the training items correctly.

Quantitative model evaluation. To get an idea whether competitive retrieval in general and CX-COM in particular explain the judgment behavior in Experiment 1, we compared it with the exemplar model, the CAM, and the blending model RulEx-J. We also included a baseline model which assumes that participants respond the same, random value in every trial. We fitted the five models to single-participant responses with a maximum likelihood estimation method. We used the BIC (Schwarz et al., 1978) to choose the best model for a single participant. A more detailed explanation about the fitting methodology, with an overview of best fitting parameter values, is presented in Appendix B.

Across different model selection criteria, the exemplar model with competitive memory retrieval and a cue-abstraction component (CX-COM) fits the data very well. It is the model with the lowest mean BIC and deviance (see Table 3) for over 60% of the participants and thus the most likely model to describe their underlying cognitive process. The second best model is the cue-abstraction model (CAM) which is the most appropriate model for approximately 25% of participants followed by RulEx-J most appropriate for just under 15% of participants. The baseline model fares far worse than the other four models in relative and in absolute terms. The average deviance for the baseline model is 200 points higher than for the other models. However, it is not always the worst model. In fact, CAM is worse than the baseline model for one participant.

A model recovery based on the design of Experiment 1 confirmed the ability of CX-COM and the CAM to discriminate when participants used these different cognitive process. When CX-COM gener-

[5] Two participants only completed eight training blocks due to technical issues.

[6] Two participants did not finish the similarity questionnaire and were excluded from the related statistics.

Table 3
Model Overview and Results of the Model Comparison

	Experiment 1				Experiment 2				
Model	Par.	Subj.	Deviance	BIC	Par.	Subj.	Deviance	BIC	Model type
CAM	5	7	1,294	1,321	6	9	1,077	1,109	Cue abstraction
Exemplar	4	0	1,375	1,397	5	2	1,119	1,146	Exemplar
CX-COM	**5**	**18**	**1,268**	**1,294**	**6**	**19**	**1,033**	**1,065**	Mixture
RulEx-J	6	4	1,285	1,317	7	1	1,075	1,112	Mixture
Baseline	2	0	1,490	1,501	2	0	1,400	1,411	—

Note. CAM = cue-abstraction model; CX-COM = competitive memory retrieval; Par. = number of parameters; BIC = mean Bayesian information criterion; Subj. = number of participants for whom a specific model had the best BIC-value. Best fitting model is shown in bold.

ated the data then 78% of the time it was identified as the data-generating model. When the CAM generated the data then 96% of the time it was identified as the data-generating model. Note that CX-COM can have overlapping predictions with the CAM, depending on the values of the attention weights. We designed Experiment 2 to discern the two models also qualitatively.

The quantitative differences in model fits are also illustrated in Figure 5 comparing participants' response distributions for each test item with the predictions of CX-COM (Figure 5A) and the second best model CAM (Figure 5B). Most test items for which a very similar training item exists (e.g., Items 3.1.1 to 1.3.3 in the figures; for reference see Table 2) show a clear peak in participants response probability for a judgment close to the value associated with the very similar training item. CX-COM is able to predict these peaks nicely while the CAM and the exemplar model cannot.

However, the BIC does not adequately reflect the functional flexibility of the models, which is particularly relevant for RulEx-J and CX-COM that appear to have a higher level of functional flexibility. Thus, we conducted a cross-validation study. The details of the cross-validation procedure are described in Appendix B and the results are shown in Appendix F. Compared with the model selection based on BIC, CX-COM still is the best model for most participants and describes 14 participants best (four participants less than previously). The mixture model RulEx-J explains 13 participants best (nine participants more than according to BIC) and CAM only explains two participants best (five participants less). Somewhat surprisingly, the results show a larger advantage for the models with a high functional flexibility, that is RulEx-J and CX-COM, compared with models that likely have a lower functional flexibility (the CAM and the exemplar model). However, there is research suggesting that cross-validation methods might favor more complex models (Browne, 2000), which might explain why RulEx-J, the arguably most complex model, had the largest gain. Overall, the result suggests that the additional functional flexibility introduced by mixture models is warranted in our task.

Discussion

In Experiment 1 we tested the CX-COM model quantitatively against competing models from the literature. CX-COM assumes that judgments are the result of (a) a competitive retrieval mechanism, and (b) a subsequent cue-abstraction mechanism that adjusts the criterion value of the recalled exemplar. We compared CX-COM with an exemplar model, a cue-abstraction model (CAM) and the blending model RulEx-J.

CX-COM captured the judgments best compared with the competing models in terms of number of assigned participants according to BIC, mean BIC, and mean deviance. It is still the best model for most participants according to the cross-validation. As Figure 5 shows, CX-COM is able to capture almost every peak in the participants' probability distribution, while the CAM, the second best model according to the BIC, cannot. Interestingly, there is no participant assigned to the (pure) exemplar model. That is, even in a multiplicative environment that is usually understood to promote the use of an exemplar process (Hoffmann et al., 2016; Juslin et al., 2008; Pachur & Olsson, 2012), we found evidence that all participants used a cue-abstraction process. This result suggests that adjustments based on beliefs about the cue–criterion relations play an important part in judgments, even when overall judgments may be best described by an exemplar model. To replicate the quantitative results and to qualitatively test the assumption that previously encountered exemplars compete for retrieval we designed Experiment 2.

Experiment 2

In Experiment 2 we focused on the prediction that sets CX-COM apart from other models in the literature: The assumption that exemplar retrieval is competitive and not integrative. For integrative retrieval, the judgment is based on the similarity to all previously encountered instances, independent of how many exemplars are similar to the probe and how strongly their criterion values differ from one another. Therefore, an integrative retrieval component predicts unimodal response distributions and no systematic variation in judgments across items.

In contrast, during competitive retrieval one exemplar is recalled on each retrieval attempt, implying that judgments for each probe vary depending on how similar (or dissimilar) the criterion values for similar exemplars in memory are. If only one exemplar in memory is highly similar to the probe, retrieval probability for this exemplar is high and it is recalled most of the time. Thus, judgments for this probe will vary only to a small extent between trials. However, if two exemplars are highly similar to the probe, both exemplars are, in principle, recalled equally often. Response variability depends on the distance between the associated criterion values. If the distance is small, say, the decision maker retrieves criterion values of 25 and 33 on a scale of 1 to 33, judgments for this probe should vary little across trials. If the distance between the associated criterion values is large, for instance nine and 33, judgments should vary strongly. In addition, the distribution of

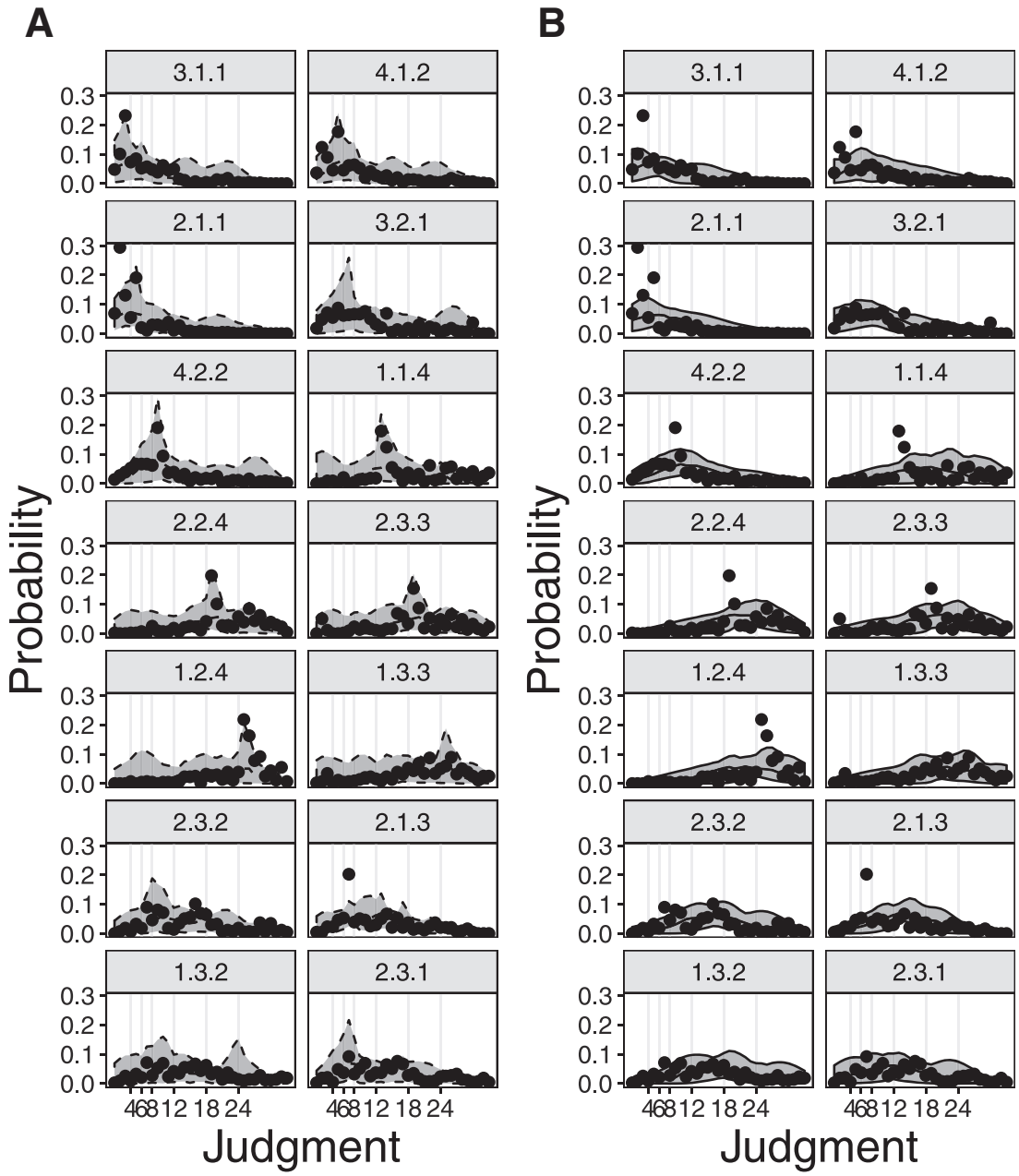


Figure 5. Percentiles .025, .5, and .975 bootstrapped, model-predicted response distributions calculated across participants for all items in Experiment 1. Participants probability to respond a specific value (between 1 and 33) is shown as dots. Light gray lines indicate criterion values of the training exemplars. (A) CX-COM-predictions (dashed line). (B) CAM-predictions (solid line).

judgments should be bimodal, within and across participants. We tested this prediction by systematically manipulating the number of exemplars with a high recall probability and the distance between associated criterion values. Recall probabilities depend on the similarity between a test item and exemplars in memory. Response variability depends on the interplay between the perceived difference in similarity and the distance in the associated criterion values.

Method

Participants. We tested 33 current or former students from the University of Basel (M_{age} = 28 years, SD = 8, range: 19–48 years).[7] The target sample size was a priori set to 30 following conventions

[7] Participant information for three participants are missing due to technical problems.

for one condition in cognitive modeling research (e.g., Hoffmann et al., 2016; Tsetsos et al., 2016). Thirty-five participants were invited through the recruitment platform of the center for Economic Psychology in Basel and 33 came at the assigned time. The experiment took on average 1 hr. Participants received course credit or an hourly payment of 20 Swiss francs. In addition, participants could earn a bonus of up to 5 Swiss francs. The study received ethics approval by the Institutional Review Board (IRB) of the Faculty of Psychology at the University of Basel.

Material. The general setup in Experiment 2 was very similar to the setup from Experiment 1. In Experiment 2 we used stimuli with four cue dimensions and three possible values on each dimension (see Figure 6). Each cue dimension was represented by a limb of a robot. Each limb had a certain number of slots for power modules (the cue values) and was associated with a different geometric form (triangle, square, circle, cross) and color (red, blue, green, brown). Participants were told to judge the overall power level of the robot (the criterion with the response scale again between 1 and 33) and that the power level depended on the number of power modules in the limbs. Figure 6 shows the training stimuli as they were presented to the participants. Positions of the different cue dimensions were partly randomized.

The overall power level was a linear function of the number of power modules in each limb,

$$j = -15 + 4 \cdot c_1 + 12 \cdot c_4, \tag{13}$$

with cue values $c_1, \ldots, c_4$. The second and the third cue were not predictive of the response.

Because the predictive cues were positively related to the criterion, participants should have been able to rapidly identify the cue–criterion relationship (see Table 4). Small values on all dimensions indicated small criterion values and large cue values indicated large criterion values respectively. However, different values on Cue Dimensions 1 and 4 required additional attention. A large value on Cue Dimension 4 and a small value on Cue Dimension 1 indicated a large criterion value, whereas a large value on Cue Dimension 1 and a small value on Cue Dimension 4 indicated a small criterion value (see Table 4 and for an example, see Figure 6).

The simple rule underlying stimulus generation allowed us to manipulate the distance between criterion values associated with training items with high recall probabilities. Thus, if the cue value of a test item on Dimension 1 was small but all other cue values were large, then a judgment based on a similar training item should have been very likely to be large as well (see test Item 1.3.3.3 in Table 4). However, if the cue value of a test item on Dimension 4 was small but all other cue values were large, a judgment based on a similar training item might be large or small (see test Item 3.3.3.1 in Table 4).

To extend the manipulation of distance between most similar exemplars conceptually to all items, we developed a measure we call the similarity neighborhood (SN; see Table 4). The SN score for a test item was calculated as the mean distance between the criterion values for those training items most similar to the test item (assuming city-block distance). The higher the mean distance is, the higher the expected variability in judgments. For example, the two training items which are most similar to test Item 3.3.3.1 are Items 3.2.2.1 (Criterion Value 9) and 3.3.3.3 (Criterion Value 33). The distance between the criterion values associated with the training items—the SN score—is 24 (see Table 4). Considering Item 1.3.3.3, the most similar items are 1.2.2.3 (Criterion Value 25) and 3.3.3.3 (Criterion Value 33) with a distance between criterion values of only eight. For items with only one most similar training item (e.g., 3.3.3.2 and 2.3.3.3) we used the mean distance between criterion values of the most and second most similar training items as SN score. Item 2.3.3.3, for example, is most similar to Item 3.3.3.3 and second most similar to Items 1.2.2.3 (with distance between criterion values of eight) and 2.3.1.2 (with distance between criterion values of 16). The SN score is, thus, 12. With this approach we considered retrieval candidates with a combined recall probability of over 90% per item ($M = 0.97$, $SD = 0.03$).

Procedure. The experiment consisted of two phases. In the training phase participants had to learn five different items by heart. They were told that they would later be asked to use these items to estimate the criterion value for novel items. During training, participants were presented with a training item on the left-hand side of the screen and had to choose the associated criterion value from the response circle on the right-hand side of the screen. Participants received feedback in all trials during training.

The maximum number of training blocks was set to 12. However, participants with 100% accuracy in three blocks (with 100% accuracy in at least two subsequent blocks) had to complete only one more block to move on to the test phase. People who failed to reach this criterion had to complete all 12 training

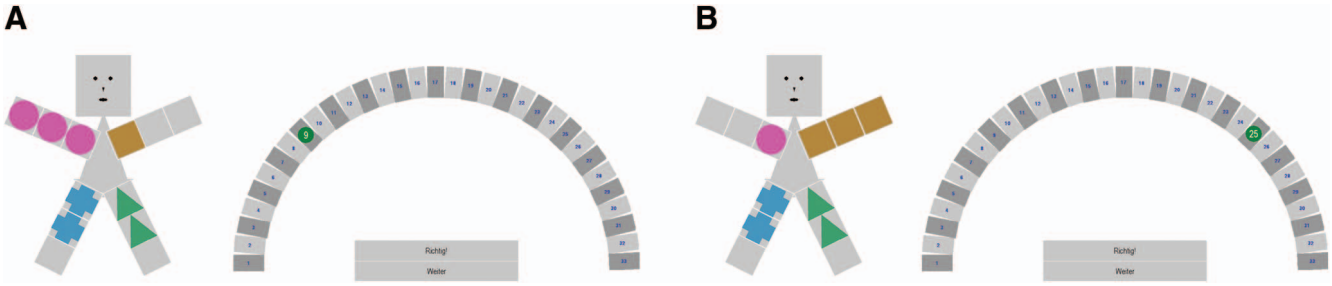


Figure 6. Example stimuli used in Experiment 2. The text on the figure is German stating "Richtig!" for "Correct!" and "Weiter" for "Next." (A) Stimulus 3.2.2.1; high value in a left extremity is associated with a low value on the scale (value 9, left part of the scale). (B) Stimulus 1.2.2.3; high value in a right extremity is associated with a high value on the scale (value 25, right part of the scale). See the online article for the color version of this figure.

Table 4
Stimuli and Results in Experiment 2

Judgment test item	Training item 1 (1.1.1.1)	9 (3.2.2.1)	17 (2.3.1.2)	25 (1.2.2.3)	33 (3.3.3.3)	SN	Results Mean (*SD*)	Results *SD* (*SD*)
1.2.2.1	**2**	**2**	4	**2**	6	16	10.08 (2.93)	3.11 (1.78)
3.2.2.3	6	**2**	4	**2**	**2**	16	26.39 (2.52)	2.77 (1.96)
2.2.2.2	4	**2**	**2**	**2**	4	11	18.22 (2.76)	3.49 (2.07)
3.3.3.1	6	**2**	4	6	**2**	24	18.04 (5.32)	4.63 (2.51)
1.3.3.3	6	6	4	**2**	**2**	8	27.34 (2.28)	2.62 (1.91)
1.1.1.3	**2**	6	4	**2**	6	24	17.52 (5.48)	3.33 (2.04)
3.1.1.1	**2**	**2**	4	6	6	8	7.93 (2.92)	2.49 (2.11)
2.1.1.1	**1**	3	3	5	7	12	6.48 (2.94)	2.5 (1.86)
1.1.1.2	**1**	5	3	3	7	20	11.3 (4.34)	3.61 (1.95)
2.3.3.3	7	5	3	3	**1**	12	29.87 (2.09)	2.25 (2.24)
3.3.3.2	7	3	3	5	**1**	20	25.59 (4.35)	4.11 (2.78)
2.2.2.1	3	**1**	3	3	5	11	11.22 (3.22)	3.04 (1.52)
3.3.2.1	5	**1**	5	5	3	24	13.98 (4.05)	3.38 (2.14)
1.2.2.2	3	3	3	**1**	5	16	16.84 (3.09)	3.7 (2.05)
1.2.3.3	5	5	3	**1**	3	8	24.78 (2.97)	2.62 (1.53)
2.2.1.2	3	3	**1**	3	5	11	15.39 (2.81)	3.3 (2.07)
2.3.2.2	5	3	**1**	3	3	11	19.99 (2.86)	3.24 (2.04)
3.2.2.1	4	**0**	4	4	4	14	11.62 (3.74)	2.31 (1.65)
2.3.1.2	4	4	**0**	4	4	12	17.47 (2.24)	2.58 (1.82)
1.2.2.3	4	4	4	**0**	4	14	23.2 (2.42)	2.94 (2.4)

Note. Distance profiles for stimuli used in Experiment 2. Items consist of four cue dimensions (cf. Figure 6). Dimension values are separated by a dot. The distances are calculated using the city-block metric and all attention weights set to 1. Items with the lowest distance are shown in bold. SN (similarity neighborhood) is the mean distance of the criterion values for similar training items. If more than one training item is most similar to a test item (first seven items), SN is the mean distance among the criterion values of these training items. If only one training item is most similar to a test item, then SN is the mean distance between the most similar and second most similar training item. Results (mean and *SD*) are calculated within participants and then averaged across participants.

blocks and then moved on to the test phase. If they failed to reach 100% accuracy in the last two blocks they were excluded from further analyses.

In the test phase, participants had to judge 20 different items without feedback. Seventeen items were unknown and three were training items (Items 3.2.2.1, 2.1.3.2, 1.2.2.3). Participants were asked to estimate the criterion values on the basis of their similarity to the items learned during training. The test phase included 10 test blocks. In each block all 20 items were presented in a randomized order, resulting in a total of 200 test trials. Participants received a bonus relative to their accuracy compared to an integrative exemplar model without cue abstraction (assuming equal weights) during the test phase. The maximal bonus was set to 5 Swiss francs.

Results

Performance. On average, participants completed training successfully after 7.9 blocks ($SD = 2.3$). Two participants failed to reach the inclusion criterion and were excluded from further analyses.

Multimodality and across-item variability. In this experiment we contrasted the competitive and integrative retrieval mechanism. Items were chosen to test two key predictions of competitive retrieval: The occurrence of multimodal response distributions within and across participants and across-item variability being a function of the similarity structure of learned exemplars. A list of all items is shown in Table 4. Recall that we hypothesized (a) that the number of training items that are very similar to a test item and their absolute difference in criterion values influences the test item's variability (e.g., Items 3.3.3.2 and 1.3.3.3 should have a lower variability than Item 3.3.3.1); and (b) that we would be able to observe a response distribution with a visible bimodal shape in test items that are very similar to two training items that have a high absolute distance between their criterion values (e.g., Item 3.3.3.1).

Multimodality: Descriptive statistics. Figure 7 shows response distributions for Items 3.3.3.1 and 1.3.3.3. The figure includes two test items that are similar to only one training item but have similar cue values (3.3.3.2 and 2.3.3.3) as well as two training items tested again during the test (a beanplot including all tested items is shown in Appendix D). For the training items, the median judgment during test corresponds to the learned criterion value and the standard deviation of responses is low. In Items 3.3.3.2 and 2.3.3.3, the standard deviation is lower than in Items 3.3.3.1 and 1.3.3.3. Additionally, the response distribution has a different shape. Especially, Item 3.3.3.1 clearly shows a multimodal response distribution with the most frequent responses close to the learned criterion values of the two most similar training items. The deviation of the most frequent responses from the learned criterion values is consistent with cue-based adjustment.

Response distributions aggregated across participants were multimodal for items with a clear effect of similarity and distance, for example, Item 3.3.3.1 in Figure 7. As an additional piece of

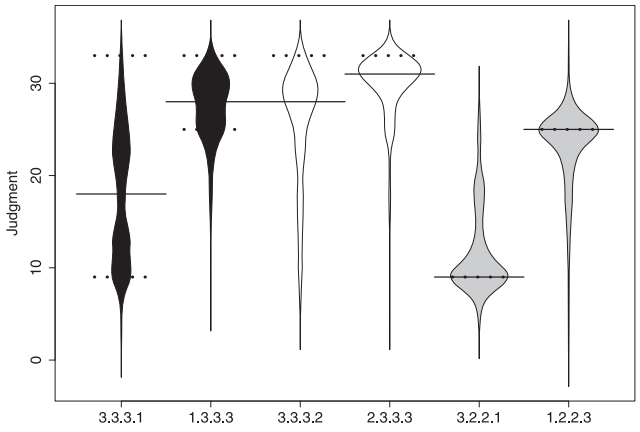


Figure 7. Participants' response distributions for items representative of the manipulation in Experiment 2. Black distributions correspond to test items with two most similar training items but a high (3.3.3.1) and low (1.3.3.3) distance between associated criterion values. White distributions correspond to test items with one most similar training item. Response distributions for training items judged at test are displayed in gray. Thick dotted lines correspond to criterion value of maximally similar training items; thin lines show the median of the distribution. The standard deviation of the kernel estimation was set to 1.06 (Scott, 1992).

evidence to corroborate our claim that this is the result of a competitive retrieval and not, for example, the result of variability in parameter settings or different strategies used between participants, we show model-predicted response distributions and observations for the two participants who were best fit by CX-COM (see Figure 8) and the CAM (see Figure 9), relative to the respective other model according to the difference in BIC. The participant best fit by CX-COM (see Figure 8) displays multimodal response distributions as predicted by CX-COM. CX-COM captures the responses of the participant well, the broad distributions predicted by the CAM do not.

Multimodality: Inferential statistics. In order to substantiate our claim and the descriptive results, we tested for multimodality across participants using Hartigan's dip test (Hartigan & Hartigan, 1985). Conceptually, Hartigan's dip test calculates the maximum distance between an empirical distribution and the best fitting unimodal distribution. In principle, all response distributions predicted by CX-COM are multimodal, the predicted response distribution can have as many modes as there are learned exemplars (cf. Figure 1). However, it can be very hard to detect this multimodality, for example, if the recall probability for one exemplar is very high. In this case, the multimodal structure of the distribution is easily confused with a distribution with long tails and a substantial number of observations are needed to classify the distribution correctly. Aggregated across participants, 17 out of 20 test items showed significant multimodality (mean $D = .06$, $p < .01$, Bonferroni corrected). The three test items which showed no signs of mulitmodality are the three training items repeated during the test phase (mean $D = 0.03$). For participants best fit by the CX-COM model (19 participants), 14 out of 20 items showed significant multimodality and for participants best fit by the CAM (nine participants) only four out of 20 items showed significant multimodality.

Across-item variability. Within participants we lacked the power to detect multimodal response distributions (see Appendix E for a post hoc power analysis). We thus tested a second hypothesis: Across-item variability is a function of the distance between criterion values of highly similar training items. The higher the distance between the criterion values of two similar training items, the more variable should responses to this test item be.

To investigate the influence of competitive exemplar retrieval on judgment variability systematically across all test items, we predicted the variability in judgments with the SN score, that is the average distance between the criterion values of similar training items. To measure variability of judgments for an item we used the standard deviation of the judgments for an item during the test phase (calculated within participant and averaged across participants), see Table 4. On average, the SN score correlated positively with the mean standard deviation, $r = .66$, $p < .01$. Because we aimed at an analysis on the participant level, we used a linear mixed effects model that predicted an item's judgment variability with its SN score as a fixed factor and items' and participants' intercepts as random factors. This model predicted the items' judgment variability significantly better than a baseline model including only participants' and items' random intercepts, $\chi^2(1) = 10.80$, $p < .01$.

Quantitative model evaluation. We fit the same models under the same conditions as in Experiment 1.[8] All models had one more free parameter because of the additional cue dimension (the respective dimension weight), except the baseline model that assumes a constant response. A detailed description of the fitted parameters and best fitting parameter values is given in Appendix B.

The CX-COM model again captured participants' judgments best. It was the most appropriate model for over 60% of the participants and had the lowest mean BIC (see Table 3). Around 30% of participants were again best described by the CAM. Surprisingly, the blending model RulEx-J fared worse than in Experiment 1 and described only one participant best according to BIC. The baseline model again fit participant responses much worse than all other models.

A model recovery with CX-COM and the CAM based on the design of Experiment 2 again supports these conclusions. When CX-COM generated the data, 94% of the time CX-COM was identified as the data-generating model. When the CAM generated the data, 87% of the time the CAM was identified as the data-generating model.

Figure 10 shows a comparison of participants' aggregated response probabilities with aggregated predictions of CX-COM (Figure 10A) and the CAM (Figure 10B). Both models describe the aggregated responses well. However, CX-COM clearly captures some peaks that the CAM cannot.

[8] Following the suggestions of anonymous reviewers we also fitted a model with competitive retrieval and without a cue-abstraction component and two possible versions of a prototype model to the data in this experiment. In one version of the prototype model we assumed the most extreme exemplars to be prototypes, i.e. 3.3.3.3 and 1.1.1.1 (Prot1). In the second version we assumed the most informative exemplars to be prototypes, i.e. 1.2.2.3 and 3.2.2.1 (Prot2). None of these three models explained any participants best and the average BIC was much higher ($BIC_{Prot1} = 1139$, $BIC_{Prot2} = 1280$, $BIC_{\text{Competitive w/o cue-abstraction}} = 1220$) than the average BIC of the CAM ($BIC = 1109$) or the CX-COM model ($BIC = 1065$).

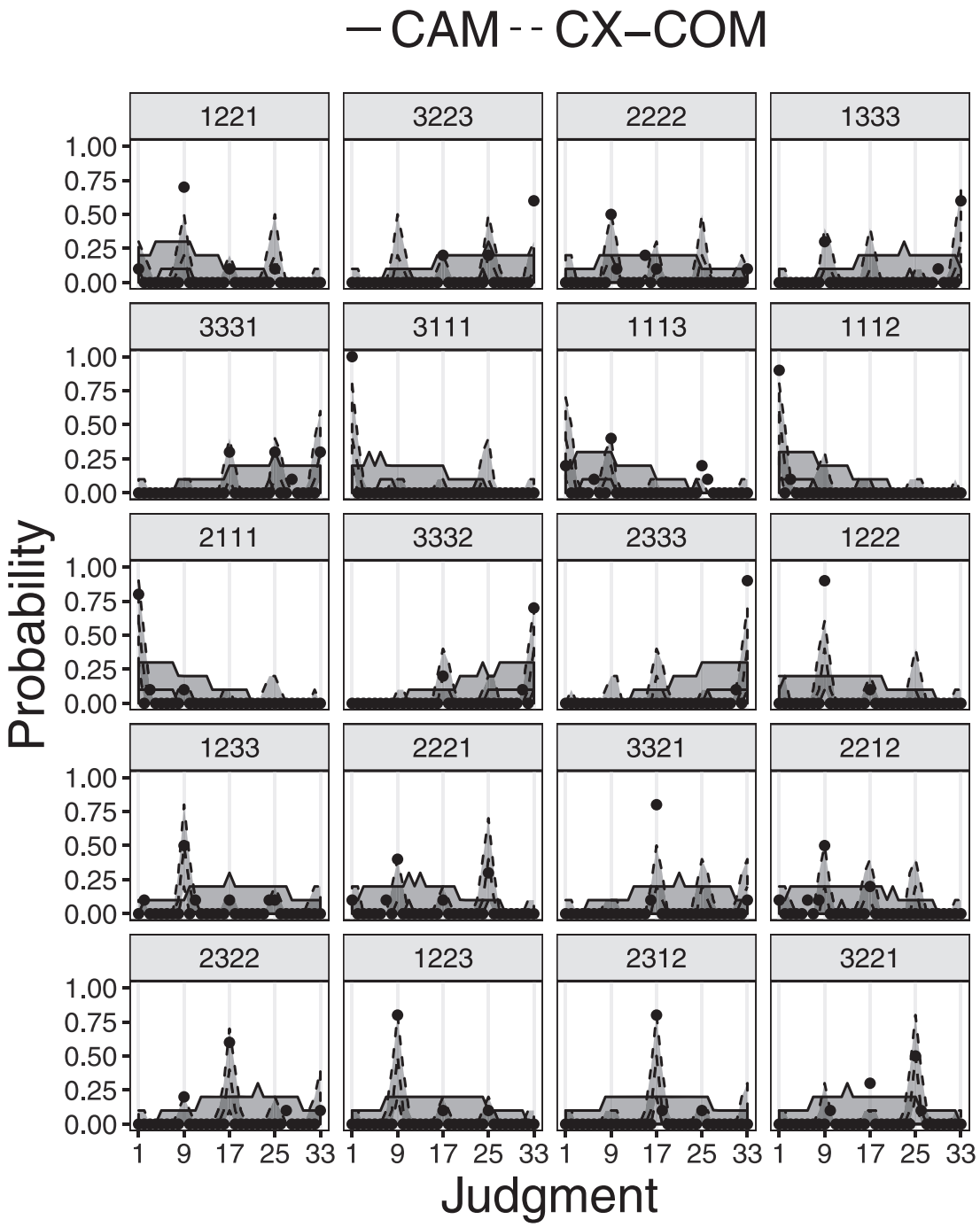


Figure 8. Percentiles .025, .5, and .975 bootstrapped, model-predicted response distributions for the participant best described by CX-COM relative to the CAM (i.e., highest difference in BICs). The participant best described by the CAM is shown in Figure 9. Light gray lines indicate criterion values of the training exemplars.

We again conducted a cross-validation in Experiment 2. The results are very similar to the results according to BIC (see Appendix F). As in Experiment 1, the two mixture models CX-COM and RulEx-J explain the responses of most participants best despite their higher functional flexibility. CX-COM is still the best model for 17 participants (than according to the BIC two participants less), whereas RulEx-J explains the judgments of 11 participants best (10 participants more than according to BIC). The number of participants best explained by the CAM drops from nine to three. The exemplar model loses all its participants. These results again suggest that the additional flexibility introduced by mixture models is warranted also according to Experiment 2.

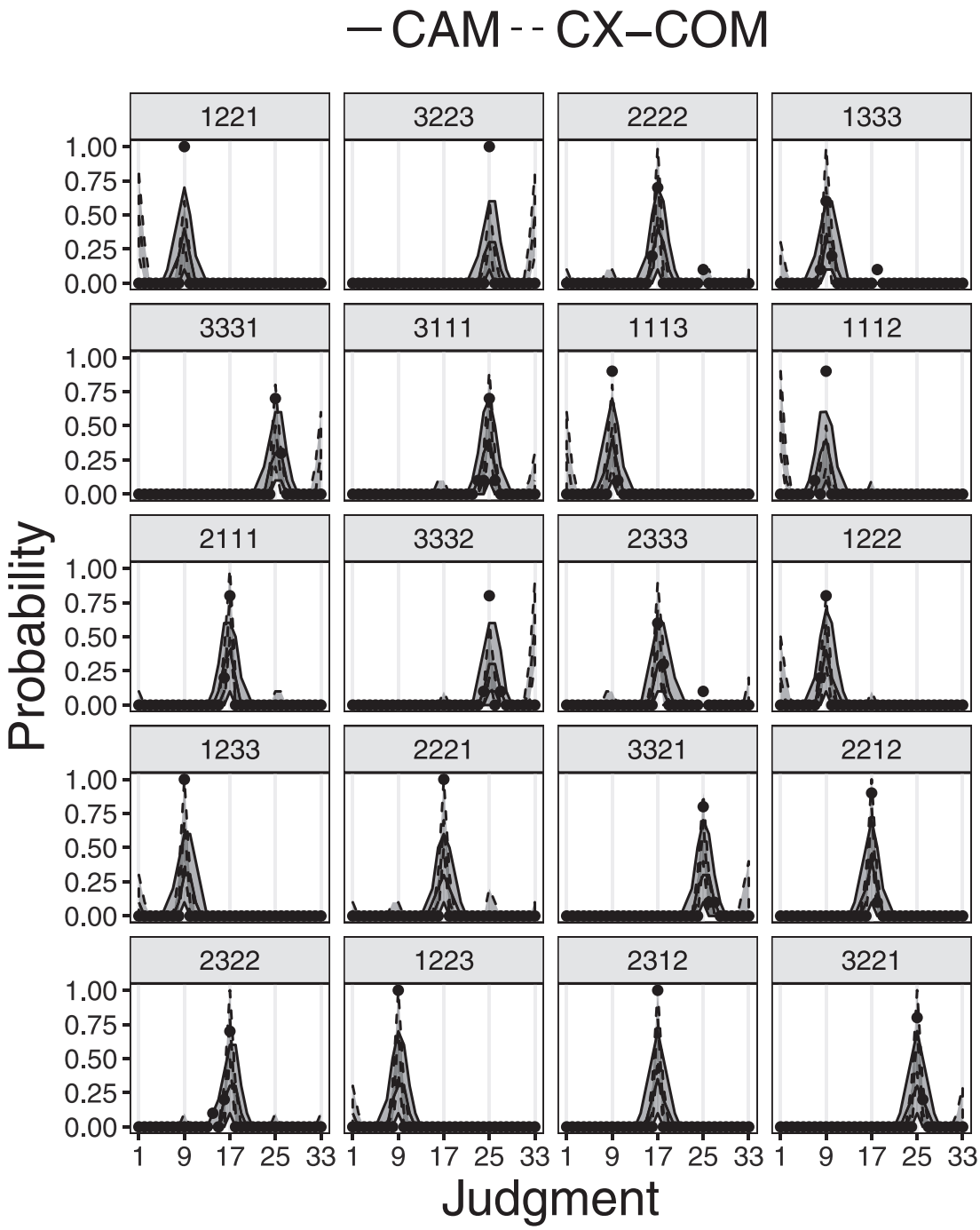


Figure 9. Percentiles .025, .5, and .975 bootstrapped, model-predicted response distributions for the participant best described by the CAM relative to CX-COM (i.e., highest difference in BICs). The participant best described by CX-COM is shown in Figure 8. Light gray lines indicate criterion values of the training exemplars.

Discussion

Experiment 2 investigated whether an exemplar model with a competitive retrieval mechanism explains judgment behavior better than the traditional exemplar model using integrative retrieval. We found that the exemplar model with an integrative retrieval mechanism could neither quantitatively nor qualitatively account for the data. Test items varied in their trial-to-trial judgment variability and most items showed clear signs of multimodality. Furthermore, judgment variability across items was a function of the distance between the criterion values associated with similar training items: A prediction that cannot be accounted for by an integrative retrieval component or the CAM. The multimodal

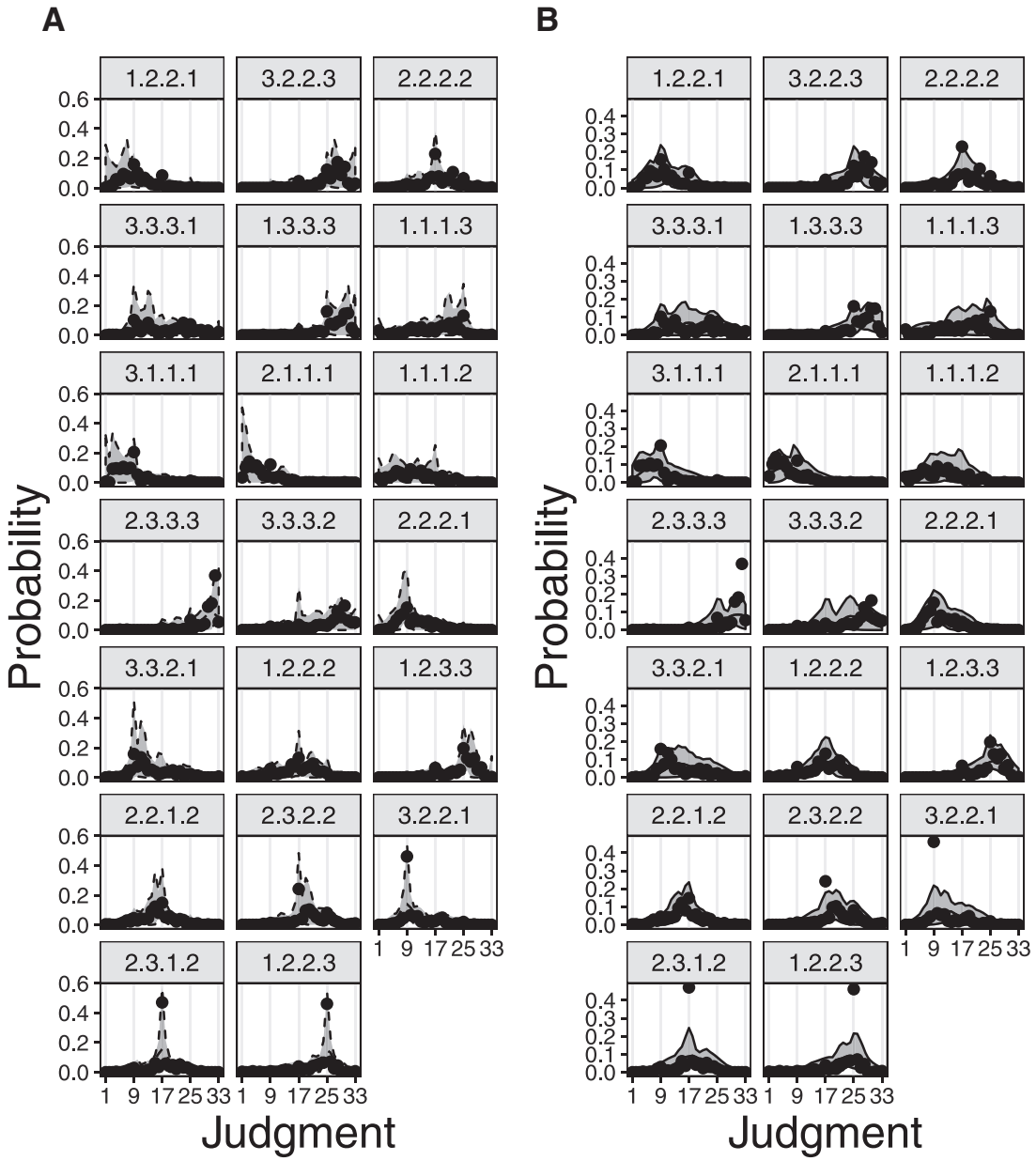


Figure 10. Percentiles .025, .5, and .975 bootstrapped, model-predicted response distributions calculated across participants for all items in Experiment 2. Participants probability to respond a specific value (between 1 and 33) is shown as dots. Light gray lines indicate criterion values of the training exemplars. (A) shows CX-COM-predictions (dashed line). (B) shows CAM-predictions (solid line).

response patterns occurred across and more importantly within participants. The existence of multimodal response distributions within participants delivers important evidence in favor of exemplar models assuming a competitive retrieval mechanism. If multimodality was found only across participants, differences in parameter settings such as attention weights or the strategies used could also explain the results.

Quantitatively, the best model was again the CX-COM model that was most appropriate for over 60% of the participants in the model comparison based on BIC. The CAM also described some participants well. One possible reason is the linear structure of the task. This type of task is known to coincide with cue-abstraction strategies (Hoffmann et al., 2016; Olsson et al., 2006). However, the CAM also assumes that judgment variability is constant across items, an assumption that was clearly violated by the majority of participants.

The results from the cross-validation likewise support CX-COM. The high number of participants best described by CX-COM and RulEx-J suggest that the majority of participants seem to rely on both cue-abstraction and exemplar processes. Still,

CX-COM clearly best captured the responses of more participants than RulEx-J. One reason for this could be that we designed it as a critical test for the prediction of multimodal responses. In sum, the results provide important evidence for competitive retrieval.

General Discussion

When people evaluate objects and situations in order to form a decision or category assignment, research suggests that knowledge of previous experiences is combined with more abstract knowledge about a specific context (e.g., Erickson & Kruschke, 1998; Juslin et al., 2008). Judgment research has largely been mute about the concrete nature of the retrieval processes and possible combinations of recalled and abstracted knowledge. The present research sought to address these two shortcomings by spelling out a new cognitive model, CX-COM. Building on established models for quantitative judgments, CX-COM introduces a competitive retrieval mechanism to describe how exemplars are activated in memory and adjusts the judgment based on the retrieved exemplar using abstracted cue knowledge. To contrast its underpinning assumptions with competing theoretical ideas, we (a) tested CX-COM quantitatively against several competitor models from the literature and (b) derived and tested a qualitative prediction about the variability and shape of response distributions induced by CX-COM's competitive retrieval mechanism. Overall, the CX-COM model was best suited to explain human judgment behavior across the two experiments. Quantitatively, the model was most appropriate for describing the data of the majority of participants and also had on average the lowest BIC values. In addition, the qualitative test supported the model's assumptions that past exemplars compete for retrieval when people make judgments (Experiment 2). We next reconsider the model's assumptions in detail and then compare the mechanisms with similar theories in judgment, categorization, and function learning research.

Competitive Retrieval From Exemplar Memory

Traditionally, exemplar models in judgment and categorization have proposed that people retrieve a composite of all previously encountered exemplars from memory. This composite does not change across trials and a constant error is assumed. This implies that judgment variability is constant across items. In contrast, in both experiments we found evidence that judgment variability systematically varied across items, indicating competitive retrieval. A stricter test of this assumption in Experiment 2 suggested that some items elicited multimodal response distributions—across and within participants. Furthermore, across-item variability was a function of an item's similarity structure, consistent with the qualitative predictions of the CX-COM model. In line with this, more participants were best described by models assuming competitive retrieval from memory in both experiments.

Taken together, these results suggest that exemplar retrieval in quantitative judgments is best described by competitive retrieval processes, corresponding to established theories on retrieval processes in episodic memory (Anderson, 1983; Logan, 2002; Ratcliff, 1978) and process-oriented versions of exemplar models (Nosofsky & Palmeri, 1997; Palmeri, 1997). The response distributions are not consistent with the assumption that judgment variability is constant across items, which is the assumption made by exemplar models with an integrative memory component, the pure CAMs, as well as blending models such as RulEx-J (Bröder et al., 2017).

Importantly, these results also highlight that the form of the response distribution and the variability of responses provide a tool to understand the nature of the involved cognitive processes (Kalish, Lewandowsky, & Kruschke, 2004). Moreover, the ability to explain and predict the expected variance in judgments is also of practical relevance given that it puts natural constraints on the expected reliability in judgments that will vary depending on the experiences of the decision maker.

Combining Cue Abstraction and Exemplar Retrieval

Although judgment research has investigated how people shift between exemplar retrieval and abstracted knowledge, little empirical work has studied the degree to which the two processes are intertwined. Within CX-COM, we assumed that cue abstraction acts on the retrieval of stored exemplars. In both experiments, CX-COM consistently outperformed an exemplar model that did not consider any cue abstraction, independently of the tested environment. This suggests that beliefs about how cues are related to the criterion influence judgments in addition to memories of similar exemplars.

Besides supporting the idea that cue-abstraction processes exist in quantitative judgments, the two experiments also provided evidence for the effects of specific exemplars. Even in the linear judgment task in Experiment 2, CX-COM described participants' judgments better than the pure CAM, although a host of research suggests that in linear tasks, judgments are usually best described by the CAMs (Hoffmann et al., 2016; Juslin et al., 2008; Pachur & Olsson, 2012). Furthermore, the multimodal response distributions follow naturally from the assumption of exemplar competition but cannot be explained by pure cue-abstraction processes.

Taken together, these results suggest that quantitative judgments are based on a combination of exemplars retrieved from memory and abstracted beliefs about the cues. They resonate well with previous empirical research showing that specific exemplars and rules simultaneously influence judgments and categorizations (Brooks & Hannah, 2006; Hahn et al., 2010; von Helversen et al., 2014) and research showing the advantage of mixture models in categorization (Erickson & Kruschke, 1998; Nosofsky et al., 1994; Vanpaemel & Storms, 2008) and function learning (DeLosh, Busemeyer, & McDaniel, 1997; Kalish et al., 2004).

Relation to Different Approaches

The two experiments provide consistent support for the CX-COM model. This new model explains how beliefs about cue–criterion relationships interact with memories about specific instances. In the following we spell out similarities and differences between CX-COM and other models and approaches in related domains.

Blending models. Blending models are based on the assumption that an exemplar and a cue-abstraction component processes information independently and the response is a weighted average of the two results. The measurement model RulEx-J (Bröder et al., 2017) is a very recent and successful implementation of this idea in the domain of multiple-cue judgments. In line with findings in

multiple-cue judgment, the mixture parameter in RulEx-J weighs the contribution of the two model components and reflects the employed strategy on the individual level and the impact of the environment or experimental instructions on the aggregate level.

In our two experiments the environments differed: In Experiment 1 we used a multiplicative environment that is known to coincide with exemplar processing and in Experiment 2 we used a linear environment that is known to coincide with cue-abstraction processes. In line with findings from the literature and especially with the results presented by Bröder, Gräf, and Kieslich (2017) we find on average a higher mixture parameter (β) in Experiment 2 (.6) than in Experiment 1 (.5; see Appendix F). These results suggest RulEx-J reflects differences in the amount of cue abstraction processes well. However, our CX-COM model outperformed RulEx-J consistently in the quantitative model comparison in both experiments. Additionally, RulEx-J is not able to account for multimodal response distributions and changes in variability across items.

Function learning. In both function learning and multiple-cue judgment, a numerical criterion has to be estimated given contextual information. However, function learning and multiple-cue judgment differ strongly in the complexity of the to-be-judged objects. In multiple-cue judgment the evaluation of objects is based on several cues with several possible values, while in function learning it is based on one numeric value. Accordingly, models from the function learning literature are rarely considered in the literature on multiple-cue judgment.

Function-learning research found that participants often extrapolated in a rule-based fashion, although they learned with single exemplars (DeLosh et al., 1997). Accordingly, theories in function learning often consider a competition between memory items (or rules) as well as mixtures between retrieval-based and cue-abstraction processes. Most notably, CX-COM could be considered as an extension of EXAM (DeLosh et al., 1997; McDaniel & Busemeyer, 2005) for judgments based on multiple cues. EXAM learns similarity-based associations between one-dimensional, quantitative inputs and outcomes. When generalizing to new patterns, it uses the distance between similar inputs to recruit a linear extrapolation mechanism. Thus, EXAM and CX-COM share the idea that a cue-abstraction mechanism adjusts the response values of a recalled response. The difference between CX-COM and EXAM mirror the different complexities of the to-be-judged objects. In the EXAM model, the cue-abstraction component considers not only the one recalled output but also two outputs with similar input values. The response is based on the proportion of change in input and output values. In contrast, CX-COM adjusts the retrieved criterion depending on the difference in cue values between the probe and the one recalled exemplar.

Knowledge partitioning. In function learning and categorization, knowledge partitioning spells out the idea that knowledge is separated into independent parcels that potentially contain mutually contradictory information (Kalish et al., 2004; Lewandowsky, Roberts, & Yang, 2006). As a result of knowledge being spread out over a space of, potentially numerical, response values in separate parcels, knowledge partitioning also predicts multimodality of responses and these patterns have been found in the domain of function learning (Kalish et al., 2004).

The most prominent model implementing knowledge partitioning in the function-learning domain is POLE (Population Of Linear Experts; Kalish et al., 2004). According to POLE, judgments are based on linear experts, that is, linear functions that are associated with each stimulus value during learning. When a new stimulus is evaluated, functions are activated based on the similarity between the new and associated stimuli. Then one rule is probabilistically selected and used to determine the response. Thus, similar to CX-COM, POLE involves a competitive selection mechanism. Consequently, an adaption of the competitive retrieval of exemplars as sketched in CX-COM is, in principle, able to explain some multimodal results POLE accounts for by assuming that sometimes an exception is recalled and adjusted according to a linear function. However, POLE stores different rules and assumes competition between these rules instead of exemplars, predicting that people also extrapolate in opposite directions depending on the exemplar they recall. CX-COM is unable to predict different extrapolation patterns on the same cue. Accordingly, although both models can predict multimodal response distributions, CX-COM and POLE differ on the items for which they predict a large variability. In CX-COM, variability is caused by training items that are activated by the same probe but differ in their criterion values; in POLE variability is caused by different functions associated with different parts of the stimulus space.

Anchoring and adjustment. On a more general level, CX-COM is also related to the idea that quantitative judgments are based on an anchoring and adjustment process (Tversky & Kahneman, 1974). According to the anchor and adjustment heuristic, people start making a judgment or an estimation by generating an initial value, the anchor. During the estimation process they then question whether the anchor provides an adequate judgment value and adjust their judgment until they are satisfied. Anchors can be internally generated values based on a memory processes or values that are externally provided in the environment (Chapman & Johnson, 2002; Epley & Gilovich, 2001; Mussweiler & Strack, 2000). Similar to internally generated anchors, CX-COM assumes that a single exemplar is retrieved from memory and the value associated with this exemplar is adjusted based on the differences in cue values and beliefs about the relation between cues and the criterion. However, CX-COM does not allow for external anchors and their influence on the judgment process. In addition, the cue-based adjustment is based on the deviation between the features of the probe and the recalled exemplar and participants' assumptions about how these features relate to the criterion but not by further knowledge.

Limitation and Future Work

We found that CX-COM accounted for judgments much better than the competing models in two experiments. However, in both experiments the number of exemplars in training was quite small and we ensured that participants memorized them very well. The open question remains of whether CX-COM still captures human judgments well if people retain more, but not necessarily intact exemplars in memory. One way people might react to more noisy exemplar representations is by giving more weight to the cue-abstraction component. Alternatively, it is possible that people will abstract prototypes or summary representations of exemplars that are clustered together. Then people might retrieve these prototypes instead of a single exemplar (Love, Medin, & Gureckis, 2004; Vanpaemel & Storms, 2008).

In both experiments we instructed participants to use the items they had learned during training to judge novel items during test (Olsson et al., 2006). We used these instructions because we aimed to provide a strong test of the retrieval assumptions underlying exemplar memory, that is, whether exemplar retrieval is integrative or competitive, and its interaction with cue knowledge. These instructions may limit the generality of CX-COM as a judgment model. However, the importance of exemplar-based processes in multiple-cue judgments studies without strategy instructions has been frequently demonstrated (Bröder & Gräf, 2018; Hoffmann et al., 2014, 2016; Juslin et al., 2008; Karlsson et al., 2007; McDaniel et al., 2018; Stillesjö, Nyberg, & Wirebring, 2019). Furthermore, a thorough, qualitative analysis of previous empirical evidence demonstrates that CX-COM can cover a broader variety of empirical findings (e.g., Juslin et al., 2008), including results that have been taken as evidence for transitions between exemplar memory and cue knowledge.

Although several memory models assume competitive retrieval, how many exemplars are recalled and combined before a response is given differs (e.g., Giguère & Love, 2013; Raaijmakers & Shiffrin, 1980). For example, the search of associative memory (SAM) model assumes that memory images resulting from a competitive retrieval process are only partly restored. To restore a full memory image, several retrievals are necessary, implying that recalled images may be a combination of values from different memory items. In this work we tested the foundation of this idea: a model that considers only one exemplar and a model that considers all exemplars to determine a response. If a subset of memory items is recalled, two scenarios are possible. If only exemplars with similar values are recalled from memory, the responses correspond to CX-COM's predictions. If very different values are recalled, they would be combined into one response similar to exemplar models with integrative retrieval. The focus on a competitive memory process in CX-COM allowed us to explore more complex and realistic memory processes in judgment research.

The present research has important implications for predictive models in the domain of quantitative judgments and evaluations. Although an integrative and a competitive retrieval mechanism as part of an exemplar model predict the same mean judgment values, the variance and actual shape of the distribution of response values might differ tremendously. Depending on the exemplars in memory that are activated in a specific context, a mean judgment value as predicted by integrative exemplar models might only be observed with a very low probability. Accordingly, the predictive power of classic exemplar models might be very low.

Conclusions

We presented a new theory and cognitive model for quantitative judgments. CX-COM models how memories about specific exemplars and general beliefs about the relation of cues with criteria are integrated into a single judgment response. Most notably, CX-COM predicts multimodal response distributions and variability in judgments based on previously encountered exemplars and the similarity of these exemplars to the item under evaluation—an aspect in judgment behavior that has been largely neglected in research. In a quantitative model comparison CX-COM consistently outperformed all competitor models. In sum, CX-COM is a promising new model of the cognitive processes underlying quantitative judgments that allows researchers to derive distinct predictions for judgment behavior in various judgment situations.

References

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior, 22,* 261–295.

Anderson, J. R., & Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin & Review, 8,* 629–647.

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,* 442.

Brehmer, B. (1994). The psychology of linear judgement models. *Acta Psychologica, 87,* 137–154.

Bröder, A., & Gräf, M. (2018). Retrieval from memory and cue complexity both trigger exemplar-based processes in judgment. *Journal of Cognitive Psychology, 30,* 406–417.

Bröder, A., Gräf, M., & Kieslich, P. J. (2017). Measuring the relative contributions of rule-based and exemplar-based processes in judgment: Validation of a simple model. *Judgment and Decision Making, 12,* 491–506.

Brooks, L. R., & Hannah, S. D. (2006). Instantiated features and the use of "rules". *Journal of Experimental Psychology: General, 135,* 133–151.

Brown, G. D., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review, 114,* 539–576.

Browne, M. W. (2000). Cross-validation methods. *Journal of mathematical psychology, 44,* 108–132.

Chapman, G. B., & Johnson, E. J. (2002). Incorporating the irrelevant: Anchors in judgments of belief and value. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 120–138). New York, NY: Cambridge University Press.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications*. Cambridge, MA: Academic Press.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 968–986.

Dougherty, M. R., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review, 106,* 180–209.

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science, 12,* 391–396.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127,* 107–140.

Garner, W. R. (2014). *The processing of information and structure*. London, UK: Psychology Press.

Giguère, G., & Love, B. C. (2013). Limits in decision making arise from limits in memory retrieval. *Proceedings of the National Academy of Sciences of the United States of America, 110,* 7613–7618.

Hahn, U., & Chater, N. (1998). Similarity and rules: Distinct? exhaustive? empirically distinguishable? *Cognition, 65,* 197–230.

Hahn, U., Prat-Sala, M., Pothos, E. M., & Brumby, D. P. (2010). Exemplar similarity and rule application. *Cognition, 114,* 1–18.

Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *The Annals of Statistics, 13,* 70–84.

Herzog, S. M., & von Helversen, B. (2018). Strategy selection versus strategy blending: A predictive perspective on single-and multi-strategy accounts in multiple-cue estimation. *Journal of Behavioral Decision Making, 31,* 233–249.

Hintzman, D. L. (1984). Minerva 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers, 16,* 96–101.

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2014). Pillars of judgment: How memory abilities affect performance in rule-based and exemplar-based judgments. *Journal of Experimental Psychology: General, 143,* 2242–2267.

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2016). Similar task features shape judgment and categorization processes. *Journal of Experimental: Psychology Learning Memory & Cognition, 42,* 1193–1217.

Juslin, P., Jones, S., Olsson, H., & Winman, A. (2003). Cue abstraction and exemplar memory in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 924–941.

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition, 106,* 259–298.

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General, 132,* 133–156.

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): A lazy algorithm for probabilistic inference from generic knowledge. *Cognitive Science, 26,* 563–607.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review, 111,* 1072–1099.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134,* 404–426.

Karlsson, L., Juslin, P., & Olsson, H. (2007). Adaptive changes between cue abstraction and exemplar memory in a multiple-cue judgment task with continuous cues. *Psychonomic Bulletin & Review, 14,* 1140–1146.

Kaufmann, E., Reips, U.-D., & Wittmann, W. W. (2013). A critical meta-analysis of lens model studies in human judgment and decision-making. *PLoS ONE, 8,* e83528.

Lewandowsky, S., & Murdock, B. B. (1989). Memory for serial order. *Psychological Review, 96,* 25–57.

Lewandowsky, S., Roberts, L., & Yang, L.-X. (2006). Knowledge partitioning in categorization: Boundary conditions. *Memory & Cognition, 34,* 1676–1688.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95,* 492–527.

Logan, G. D. (2002). An instance theory of attention and memory. *Psychological Review, 109,* 376–400.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: A network model of category learning. *Psychological Review, 111,* 309–332.

Macrae, C., Bodenhausen, G. V., Milne, A. B., Castelli, L., Schloerscheidt, A. M., & Greco, S. (1998). On activating exemplars. *Journal of Experimental Social Psychology, 34,* 330–354.

Mata, R., von Helversen, B., Karlsson, L., & Cüpper, L. (2012). Adult age differences in categorization and multiple-cue judgment. *Developmental Psychology, 48,* 1188–1201.

McDaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: Comparison of rule-based and associative-based models. *Psychonomic Bulletin & Review, 12,* 24–42.

McDaniel, M. A., Cahill, M. J., Frey, R. F., Rauch, M., Doele, J., Ruvolo, D., & Daschbach, M. M. (2018). Individual differences in learning exemplars versus abstracting rules: Associations with exam performance in college science. *Journal of Applied Research in Memory & Cognition, 7,* 241–251.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238.

Mussweiler, T., & Strack, F. (2000). The use of category and exemplar knowledge in the solution of anchoring tasks. *Journal of personality and social psychology, 78,* 1038–1052.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 104–114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–61.

Nosofsky, R. M. (1997). An exemplar-based random-walk model of speeded categorization and absolute judgment. *Psychological Review, 104,* 266–300.

Nosofsky, R. M. (2014). The generalized context model: An exemplar model of classification. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 18–39). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511921322.002

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104,* 266–300.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101,* 53–79.

Olsson, A.-C., Enkvist, T., & Juslin, P. (2006). Go with the flow: How to master a nonlinear multiple-cue judgment task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 1371–1384.

Pachur, T., & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology, 65,* 207–240.

Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 324–354.

Palmeri, T. J., Wong, A. C., & Gauthier, I. (2004). Computational approaches to the development of perceptual expertise. *Trends in Cognitive Sciences, 8,* 378–386.

Platzer, C., & Bröder, A. (2012). Most people do not ignore salient invalid cues in memory-based decisions. *Psychonomic Bulletin & Review, 19,* 654–661.

Pleskac, T. J., Dougherty, M. R., Rivadeneira, A. W., & Wallsten, T. S. (2009). Random error in judgment: The contribution of encoding and retrieval processes. *Journal of Memory and Language, 60,* 165–179.

Raaijmakers, J. G., & Shiffrin, R. M. (1980). Sam: A theory of probabilistic search of associative memory. *Psychology of Learning and Motivation, 14,* 207–262.

Raaijmakers, J. G., & Shiffrin, R. M. (1992). Models for recall and recognition. *Annual Review of Psychology, 43,* 205–234.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rouder, J. N., & Ratcliff, R. (2006). Comparing exemplar-and rule-based theories of categorization. *Current Directions in Psychological Science, 15,* 9–13.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. New York, NY: Wiley and Sons.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of mathematical psychology, 1,* 54–87.

Stillesjö, S., Nyberg, L., & Wirebring, L. K. (2019). Building memory representations for exemplar-based judgment: A role for ventral precuneus. *Frontiers in Human Neuroscience*. Advance online publication. http://dx.doi.org/10.3389/fnhum.2019.00228

Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M., & Summerfield, C. (2016). Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences of the United States of America*. Advance online publication. http://dx.doi.org/10.1073/pnas.1519157113

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review, 15,* 732–749.

von Helversen, B., Herzog, S. M., & Rieskamp, J. (2014). Haunted by a doppelgänger: Irrelevant facial similarity affects rule-based judgments. *Experimental psychology, 61,* 12–22.

von Helversen, B., Mata, R., & Olsson, H. (2010). Do children profit from looking beyond looks? From similarity-based to cue abstraction processes in multiple-cue judgment. *Developmental Psychology, 46,* 220–229.

von Helversen, B., & Rieskamp, J. (2008). The mapping model: A cognitive theory of quantitative estimation. *Journal of Experimental Psychology: General, 137,* 73–96.

von Helversen, B., & Rieskamp, J. (2009). Models of quantitative estimations: Rule-based and exemplar-based processes compared. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 867–889.

Yang, L.-X., & Lewandowsky, S. (2004). Knowledge partitioning in categorization: Constraints on exemplar models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 1045–1064.

Appendix A

Reanalysis of Previous Data and Its Limitations

We reanalyzed data from Hoffmann et al. (2014, 2016) who systematically investigated judgment strategies across different environments without strategy instructions. For all the different environments we fitted the CX-COM model, the CAM, and the exemplar model. In the environments from the 2016 article, CX-COM explains most participants best in the one-dimensional linear environment (22 out of 32 in the first and 24 out of 32 in the second variant) and in the multi-dimensional multiplicative environment (all 32), and it explains about half of the participants best in the multidimensional quadratic environment (16 out of 32). In the multidimensional linear environment, the CAM is the best model with 18 out of 32 and CX-COM explains only seven participants best. In the environments tested in the 2014 article, CX-COM is the best model in the multiplicative condition (267 out of 287) and the CAM is the best model in the linear condition (176 out of 287).

Unfortunately, these environments were not designed for tearing apart CX-COM from other models of human judgment. A model recovery suggested that CX-COM and the CAM could not be distinguished because CX-COM can make similar predictions as a cue-abstraction model. Importantly, CX-COM and the CAM often only differ when predicting full response distributions instead of average responses. Therefore, it is necessary to observe many responses on the same test items to successfully recover CX-COM and contrast it with a CAM. Previous studies on judgment research, however, usually tested judgments for many test items but did not assess full response distributions for single items. This data structure thus poses a problem for evaluating CX-COM's performance using previously published data. In the multidimensional linear environment from the 2016 paper, for example, CX-COM's recovery rate is around 50% while CAM's is around 90%. In contrast, in the multidimensional multiplicative condition, the CAM can only be recovered in less than 40% while CX-COM can be recovered in more than 90% of all cases.

(*Appendices continue*)

Appendix B

Fitting and Implementation Details

All analyses are done with the R programming language (R Core Team, 2015).

For each model–participant combination we searched for the best fitting parameter setting by minimizing the models' negative log-likelihood. To find the best fitting model for every participant we used the Bayesian information criterion (BIC; Schwarz et al., 1978) to penalize more complex models.

All exemplar models (or model components) were fit to the data with the city-block distance (see Equation 3, with $r = 1$). Both exemplar and cue-abstraction models contain parameters reflecting the importance/attention given to the cue dimensions. To be able to fit both processes simultaneously we estimated only one set of dimension weights for both the exemplar retrieval process and the cue-based adjustment, so $w_i = b_i$ for each cue-dimension i. To do this, we freely estimated one weight for each cue dimension. In the cue-abstraction component, the adjustment was calculated according to the estimated weights. In the exemplar components we determined the attention weights w_i and the sensitivity parameter c by setting it to the sum of the absolute weights. We then calculated the attention weights by dividing the absolute weights of the respective dimensions by c. Hence, the attention weights in the exemplar process varied between 0 and 1 and summed up to 1 following the constraints usually assumed in exemplar models. Best fitting parameter values for both experiments are shown in Appendix F.

For parameter estimation we used a combination of grid search and nonlinear optimization. The grids had a step size of 1 and the overall size of the grid was informed by the true parameters of the functions underlying stimulus creation. In Experiment 1 the borders of the grid were set to -10 and 10 and in Experiment 2 to -20 and 15. These choices yielded 21 optimization searches for each model in Experiment 1 and 46 in Experiment 2. For each search, the starting parameter values were set to a random value between two subsequent grid values. Starting values outside the range of possible parameter values were ignored and set to the respective borders of the range instead. As optimization algorithm we used the "nlminb" function in the "stats" package of the R programming language (R Core Team, 2015).

Cross Validation

Although CX-COM possesses the same number of parameters as the CAM and only one parameter more than the exemplar model, CX-COM may be more prone to overfitting than the CAM or the exemplar model. CX-COM is a mixture model which functionally traverses between pure exemplar and pure cue-abstraction predictions. Thus, similar to the mixture model RulEx-J Bröder et al. (2017), its functional complexity likely exceeds the functional complexity of pure exemplar and cue-abstraction models. To better understand whether the functional complexity is warranted given the data we conducted a cross validation.

For each participant, we split the set of observations for each test item into half and randomly assigned one half to the training set and the other half to the validation set. In Experiment 1, this resulted in splitting the observations into one set with seven observations and one set with eight observations. In Experiment 2, this resulted in two sets with five observations each. We then estimated the model parameters for each model and participant using participants' responses to items in the training set and predicted the responses from the validation set (and vice versa). Reported results are the mean of these two predictions (see Appendix F).

Appendix C

Similarities

In Experiment 1 participants rated how similar every training item was to every test item. Appendix G shows the aggregated results. The aggregated similarity ratings corresponded to the aggregated similarities predicted by the CX-COM model for all participants ($r = .89$, $p < .01$). An analysis on the individual level confirms the results on the aggregate level (mean $r = .70$, $p < .01$) with every individual correlation being significant (individual p values were corrected for multiple comparisons using the Bonferroni correction method).

(*Appendices continue*)

Appendix D

Beanplot for All Items Tested in Experiment 2

Figure D1 shows a beanplot for all items in Experiment 2.

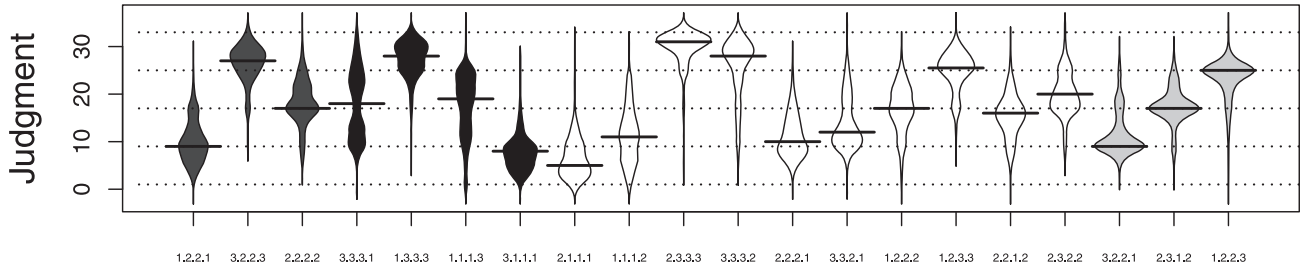


Figure D1. Participants' response distributions (as shown in Figure 7) for all items tested in Experiment 2. Dark gray distributions correspond to items with three most similar training items, black distributions to test items with two most similar training items, white distributions to item with one most similar training item, and light gray distributions correspond to training items repeated during the test phase. Thick dotted lines correspond to the criterion values of the training items; thin lines show the median of the distribution. The standard deviation of the kernel estimation was set to 1.06 (Scott, 1992).

Appendix E

Post-Hoc Power Analysis of Multimodality Within Participant and Item in Experiment 2

Within participants we only have 10 observations per item in Experiment 2 and lacked the power to detect multimodality. Out of the 620 statistical tests on the participant/item level, approximately 20% were significant ($p < .05$, not corrected). To better understand how many observations would have been needed to have enough power, we performed a post-hoc power analysis. Thereby we utilized the fact that CX-COM predicts a multimodal response distribution. We drew 10, 20, 50, 100, and 1,000 samples (with 100 repetitions each) from the response distributions predicted by CX-COM for each participant/item combination. In the case of 10 samples, multimodality was only detected in 17% of all tests, followed by 21%, 53%, 82%, and 100% in the case of 1,000 samples.

To test how often false positive results occur, we checked how likely a normal distribution is falsely identified as being multimodal by the dip test: Drawing 10 samples (same number as observations per participant and item as in the experiment) with 1,000 repetitions from normal distributions with variances of 1, 2, and 5 there were less than 0.2% false positive results in all three cases.

(*Appendices continue*)

Appendix F

Model Fit Parameters

Parameter	CAM	Exemplar	CX-COM	RulEx-J	Baseline
			Experiment 1		
$w_1 = b_1$ (Dimension weight)	−.89	3.12	.12	35.95	—
$w_2 = b_2$ (Dimension weight)	3.79	1.94	.67	−3.99	—
$w_3 = b_3$ (Dimension weight)	2.89	6.35	.79	−30.97	—
k (Intercept)	1.79	—	—	1468.69	—
c (Sensitivity)	—	11.42	1.58	1	—
α (Cue-based adjustment)	—	—	−101.39*	—	—
β (Model selection probability)	—	—	—	.5	—
σ^2 (Error variance)	5.45	6.51	2.93	5.33	8.51
Mean in baseline model	—	—	—	—	14.58
Mean deviance	1,294	1,375	1,268	1,285	1,490
Mean BIC	1,321	1,397	1,294	1,317	1,501
Number of parameters	5	4	5	6	2
Number of best fitted participants	7	0	18	4	0
Mean deviance (CV fit)	645	687	679	643	—
Mean deviance (CV prediction)	654	691	647	650	—
Number of best predicted participants (CV)	2	0	14	13	—
			Experiment 2		
$w_1 = b_1$ (Dimension weight)	6.71	.98	.99	4.16	—
$w_2 = b_2$ (Dimension weight)	2.43	.45	.27	1.5	—
$w_3 = b_3$ (Dimension weight)	2.56	.92	.43	1.89	—
$w_4 = b_4$ (Dimension weight)	1.58	.42	.42	.96	—
k (Intercept)	−8.7	—	—	66.97	—
c (Sensitivity)	—	2.76	2.11	8.5	—
α (Cue-based adjustment)	—	—	33.64*	—	—
β (Model selection probability)	—	—	—	.6	—
σ^2 (Error variance)	3.80	4.2	2.02	3.78	8.06
Mean in baseline model	—	—	—	—	17.26
Mean deviance	1,077	1,119	1,033	1,075	1,400
Mean BIC	1,109	1,146	1,065	1,112	1,411
Number of parameters	6	5	6	7	2
Number of best fitted participants	9	2	19	1	0
Mean deviance (CV fit)	535	558	518	536	—
Mean deviance (CV prediction)	549	565	527	546	—
Number of best predicted participants (CV)	1	0	17	11	—

Note. CAM = cue-abstraction model; CX-COM = competitive memory retrieval. Mean parameter values, Bayesian information criterion (BIC), and model descriptions. BIC and number of participants of the best fitting model are marked in bold. The mean deviance of the cross validation (CV) are averaged across the two cross-validation sets (see Appendix B) for fits (CV fits) and predictions (CV predictions), number of best-predicted participants according to cross validation (CV).

* The high value for α in both experiments stems from a small number of participants with α values above 100. These participants were poorly fit by CX-COM and were not included in participants best fit by the model. The median for α is 2.86 in Experiment 1 and 3.17 in Experiment 2. The mean over participants best fit by the CX-COM model is 1.05 in Experiment 1 and 6.92 in Experiment 2.

(*Appendices continue*)

Appendix G

Similarity Ratings in Experiment 1

Judgment test items	Training item 4 (3.1.2)	6 (2.1.2)	8 (3.2.2)	12 (1.1.3)	18 (2.2.3)	24 (1.2.3)
3.1.1	**7.00** (2.68)	4.52 (2.71)	6.14 (2.20)	2.71 (2.70)	2.19 (1.78)	1.43 (1.25)
4.1.2	**8.19** (1.47)	4.67 (2.82)	5.67 (2.76)	2.05 (1.75)	2.14 (1.53)	1.33 (1.56)
2.1.1	4.76 (2.64)	**6.90** (2.36)	4.76 (2.36)	2.33 (1.85)	2.05 (1.83)	1.24 (1.45)
3.2.1	5.67 (2.87)	3.38 (2.71)	**7.67** (1.93)	1.62 (1.69)	2.67 (1.80)	2.48 (2.20)
4.2.2	6.38 (2.75)	3.76 (2.41)	**7.95** (1.53)	1.62 (1.40)	2.90 (2.10)	1.90 (1.79)
1.1.4	2.29 (1.68)	2.29 (1.85)	1.67 (1.49)	**7.86** (1.82)	5.29 (1.95)	5.29 (2.24)
2.2.4	1.90 (1.84)	2.62 (2.25)	2.67 (2.13)	5.33 (2.33)	**8.14** (1.31)	6.05 (2.48)
2.3.3	2.10 (2.07)	2.90 (2.23)	4.14 (2.50)	3.71 (2.59)	**7.57** (1.29)	5.14 (2.92)
1.2.4	1.71 (2.03)	1.76 (1.61)	2.43 (1.43)	6.29 (1.65)	6.67 (1.71)	**8.24** (1.51)
1.3.3	1.76 (1.45)	1.86 (1.62)	2.67 (1.91)	4.52 (1.86)	5.86 (2.26)	**7.24** (2.23)
2.3.2	2.81 (2.14)	5.05 (2.65)	4.33 (2.54)	2.10 (2.26)	5.00 (2.26)	2.76 (2.41)
2.1.3	5.38 (2.69)	6.62 (2.87)	4.00 (2.59)	6.71 (2.45)	6.95 (2.01)	4.62 (2.50)
1.3.2	3.62 (2.25)	2.67 (1.77)	3.57 (2.25)	3.00 (1.70)	3.95 (2.16)	5.43 (2.56)
2.3.2	1.76 (1.41)	2.76 (1.79)	3.67 (2.13)	1.43 (1.91)	3.76 (2.07)	2.76 (2.30)

Note. Participants' mean similarity ratings (and standard deviations) in Experiment 1. Highest perceived similarity is marked in bold.

Received August 14, 2018
Revision received August 23, 2019
Accepted August 27, 2019 ■

The Similarity-Updating Model of Probability Judgment and Belief Revision

Rebecca Albrecht[a*], Mirjam A. Jenny[a,b*], Håkan Nilsson[a,c], and Jörg Rieskamp[a]

[a]University of Basel

[b]Max Planck Institute for Human Development

[c]Uppsala University

Author note

* These authors contributed equally to this work. This research was supported by grants from the Swiss National Science Foundation (100014_126721/1 and 100014_138174/1) to JR.
Correspondence concerning this article should be addressed to Mirjam Jenny, Harding Center for Risk Literacy, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. E-mail: jenny@mpib-berlin.mpg.de. A poster with the title “How irrelevant information influences people's probability judgments” has been presented by Mirjam Jenny at the 33rd Annual Meeting of the Society for Judgment and Decision Making (Nov 16-19, 2012, Minneapolis) detailing a previous version of the cognitive model and results for the first three studies.

Abstract

When people revise their beliefs on the basis of new information, they sometimes take nondiagnostic information into account. Although this "dilution effect" has been found in diverse areas, few studies have modeled the underlying cognitive processes. To explain the cognitive processes we suggest a similarity-updating model, which incorporates a similarity judgment inspired by similarity models of categorization research and a weighting-and-adding process with an adjustment mechanism suggested in judgment studies. We illustrate the predictive accuracy of the model for probability judgments and belief revision with four experimental studies. Participants received samples from two options and had to judge to which option the samples belonged. The similarity-updating model predicts that this probability judgment is a function of the similarity of the sample to the options. When presented with a new sample, the previous probability judgment is updated with a second probability judgment by taking a weighted average of the two and adjusting the result according to a confirmation-based adjustment mechanism. The model describes people's probability judgments well and outcompetes a Bayesian cognitive model and an alternative probability-theory-plus-noise model. The similarity-updating model accounts for several empirical results, such as the dilution effect, according to which nondiagnostic information "dilutes" the probability judgment, order effects, and the finding that probability judgments are invariant to sample size. In sum, the similarity-updating model provides a plausible account of human probability judgment and belief revision.

Keywords: probability judgment, belief updating, similarity, dilution effect, cognitive modeling

The Similarity-Updating Model of Probability Judgment and Belief Revision

Judging probabilities correctly and making good decisions under uncertainty are crucial skills and can affect any area of our lives, be it, for example, finance, education, or health care. What is the probability that my newly acquired stocks will rise within the next month? What is the probability that my child will get into an Ivy League university? And what is the probability that my mother will recover from her hip injury within the next year? People seldom make such probability judgments based on only one piece of data. Rather, multiple pieces of information are usually acquired sequentially, possibly with a lag of a few minutes, hours, days, or weeks. Therefore, initially formed probability judgments have to be dynamically revised in light of new information. To investigate the cognitive processes underlying such judgments we used the classic "book bags and poker chips" task (Edwards, 1968), with two card decks consisting of three different colors and several sequentially drawn card samples. In this article we introduce the similarity-updating model, with which we describe the cognitive process of how people make and revise probability judgments. This model can explain when and why people's judgments deviate from probability theory.

Probability Judgments and Belief Revision

Probability judgments go beyond simple binary predictions as to whether a certain event will happen. Making a probability judgment means assigning a degree of belief in the occurrence of the event in question. How should people ideally deal with uncertainty and how should probability judgments be formed? Mathematics prescribes probability theory to compute probabilities. Specifically, Bayesian theory prescribes how to compute probabilities and likelihoods and how to update them in light of new information. Research has shown that people's probability judgments often do not follow the rules of probability theory in various situations. Well-described phenomena in the probability judgment literature, such as conservatism (Dougherty et al., 1999; Edwards, 1968), base-rate neglect (e.g., Bar-Hillel,

1980), sub- and superadditivity (Dougherty & Hunter, 2003; Macchi et al., 1999; Tversky & Koehler, 1994), and the conjunction fallacy (e.g., Tversky & Kahneman, 1983), exemplify that people's judgments are often inconsistent with probability theory. The dilution effect (e.g., Nisbett et al., 1981) describes how people's probability judgments are influenced by nondiagnostic information. In consecutive probability judgments, although a second piece of evidence supports different hypotheses equally, people still decrease their initial probability judgments. This dilution effect is central to the present work. We propose a new model that combines a set of previously discussed theories to explain this behavior. But first, we focus on belief revision (i.e., on how people revise or update their probability judgments in light of new information) and discuss the dilution effect as well as other important findings that characterize probability judgments.

Characteristics of Probability Judgments

People show certain qualitative effects in their probability judgments that often cannot be explained by normative models of probability assessment such as Bayesian probability theory. Next, we discuss important qualitative effects of probability judgments in more detail: the dilution effect, the confirmation effect, order effects, converging evidence, insensitivity to sample size, and regression to the mean.

The Dilution Effect

How people update their beliefs has been investigated in various ways. In dilution tasks, participants typically have to judge the probability of event X based on a diagnostic information unit E_1: $p(X \mid E_1)$. Next, they receive a nondiagnostic piece of information, E_2, and judge $p(X \mid E_1, E_2)$. Bayes's theory postulates that because E_2 is nondiagnostic, $p(X \mid E_1, E_2)$ should equal $p(X \mid E_1)$. The dilution effect is thus said to be observed if $p(X \mid E_1)$ is greater than $p(X \mid E_1, E_2)$.

In Shanteau's (1975) classic numerical dilution study, participants observed samples of red and white beads being drawn with replacement from either a box consisting of 70 white and 30 red beads or a box with 30 white and 70 red beads. After the presentation of each sample, participants had to estimate the probability that the sample came from the 70/30 box by sliding a pointer along an unmarked ruler. Participants' mean probability judgments decreased after the presentation of a nondiagnostic sample. In contrast, when a nondiagnostic sample followed an already nondiagnostic sample, the average probability judgment remained the same. These findings held whether or not the ruler was put to the center of the scale after every judgment and whether or not participants responded on a log odds scale. In our studies reported below, we used a similar task.

LaBella and Koehler (2004) demonstrated that when people base their judgments of the probabilities of sampling marbles out of urns on a mix of diagnostic and nondiagnostic samples in aggregated samples, the resulting probability judgments are less extreme than if they based their probability judgments on diagnostic samples alone. Thus, nondiagnostic samples dilute people's probability judgments.

The influence of nondiagnostic information on judgments has been found in several everyday-life scenarios: For instance, in legal decision making, confidence in a verdict increased after an independent verdict was received (McKenzie et al., 2002). In social reasoning, students' estimates about other students changed after they received nondiagnostic information (Peters & Rothbart, 2000). In auditor judgments, inexperienced auditors were affected by nondiagnostic information (Shelton, 1999) and auditors in general underweighted diagnostic information (Waller & Zimbelman, 2003).

In sum, the dilution effect has been observed in various areas (Macrae et al., 1992; McKenzie et al. 2002; Meyvis & Janiszewski, 2002; Peters, Dieckmann et al., 2007; Shanteau, 1975). In the present work we suggest a psychologically plausible model of why

the dilution effect occurs and thereby make a novel contribution to the various research areas in which it is studied.

Factors Mitigating the Dilution Effect

Dilution effects do not always occur. For instance, although auditor judgments by relatively less experienced auditors are diluted by nondiagnostic information, judgments of experienced auditors are not (Shelton, 1999). In legal decision making, people are relatively immune to the dilution effect when they have a lot of diagnostic information (Smith et al., 1999). In perceptual decision making, the dilution effect occurs less frequently and/or less strongly if naturalistic stimuli that promote automatic processing are used (Hotaling et al., 2015). In social reasoning, whether the dilution effect occurs depends on the typicality of the nondiagnostic information. Testing boundary conditions for the dilution effect in verbal tasks, Peters and Rothbart (2000) showed that the typicality of nondiagnostic information determines whether subsequent judgments are affected. When participants predicted the number of books a fraternity member would read outside of class assignments, whether nondiagnostic information diluted their initial judgment depended on whether this information increased (does not like parties), decreased (is extroverted), or did not change (juggles) the typicality of the target person as a fraternity member. The dilution effect occurred only after the presentation of information that was atypical (does not like parties) for the target category (fraternity member).

Sanborn et al. (2020) presented evidence thought to indicate that the dilution effect cannot be ascribed to an inaccurate combination of diagnostic and nondiagnostic information but rather to an overestimation of the diagnostic information alone. Specifically, in their task, diagnostic evidence included only one source of information (e.g., the type of ice cream) while additional, nondiagnostic evidence was a combination of two sources, the diagnostic bit of information plus a nondiagnostic bit of information (the type of ice cream plus the type

of cone). They explained their results with a theory based on missing features, assuming that participants filled in information based on the most likely hypothesis (in their example the shop the ice cream was bought from). Their theory cannot be used to explain the dilution effect in our setting because there are no missing features in the type of task we used (book bags and poker chips).

Psychological Models That Explain the Dilution Effect

Because they did not spell out the cognitive processes that cause people to adjust their probability judgments on the basis of nondiagnostic information, the previously discussed empirical studies provide few theoretical explanations for the dilution effect. Although it has been discussed that averaging processes might produce the dilution effect (Shanteau, 1975; Troutman & Shanteau, 1977; but see LaBella & Koehler, 2004), these models have not been tested within a cognitive modeling framework. Alternative theoretical explanations such as expectancy and representativeness have been discussed (Tetlock & Boettger, 1989; Troutman & Shanteau, 1977). Expectancy describes the idea that upon sampling a first sample, people form hypotheses about how subsequent samples are likely to look. If these expectations are not met by the nondiagnostic sample, people become less certain about which of the two options is the one producing the samples. This account predicts that a first sample that is nondiagnostic does not have any influence on subsequent samples because it does not elicit a directed hypothesis. Nondiagnostic first samples have been found to influence subsequent judgments, however, which speaks against this account (Troutman & Shanteau, 1977). According to the representativeness account, people predict outcomes that are representative of an option. Therefore, similarity between an option and an outcome depends on the number of common features (Tetlock & Boettger, 1989). The representativeness account has been tested in a cognitive modeling framework (representativeness as prototype similarity and

representativeness as relative likelihood; Nilsson et al., 2005) and we further test this account in the present article.

Verbal similarity theories (e.g., Nisbett et al., 1981; but see Peters & Rothbart, 2000) have provided yet another theoretical explanation for the dilution effect. According to Nisbett et al. (1981), a nondiagnostic piece of evidence decreases the perceived similarity between a hypothesis and previously presented evidence, which dilutes belief in the hypothesis, instead of an averaging process diluting belief. We implemented these ideas in a computational model, the similarity-updating model, which can explain the dilution effect and delivers clear testable predictions on a trial-by-trial level.

The Confirmation Effect

People often have to adjust their probability judgments when sequentially encountering pieces of information (belief updating is discussed in more detail below). Often people seek out information to confirm an earlier hypothesis instead of information that might reject it (Jones & Sugden 2001; Wason, 1968). Similarly, people tend to interpret information in a way that confirms their hypotheses (Lord et al., 1979; Plous, 1991). For example, the predecisional distortion of information effect (Russo et al., 1998) describes how people tend to interpret new, nondiagnostic information as being in favor of their original hypothesis if they are asked repeatedly for their preferences (stepwise evolution of preference paradigm; for an overview see Russo, 2015).

The confirmation effect, especially in the form of predecisional distortion of information, can be seen as the opposite of the dilution effect. LaBella and Koehler (2004) investigated in detail in what situations people show a dilution effect or a confirmation effect. They observed the dilution effect when they compared probability judgments that were based on a mix of diagnostic and nondiagnostic information in an aggregated sample rather than on diagnostic information alone. But when they had participants revise their initial probability

judgments that were based on diagnostic information after the presentation of additional nondiagnostic information, they did not observe the dilution effect. More specifically, people showed no dilution effect on average when a diagnostic sample was followed by a neutral sample and they showed the reverse effect, namely, an increased probability judgment (confirmation effect), when a diagnostic sample was followed by a mixed nondiagnostic one. According to Bayesian theory, the type of nondiagnostic information should not play a role (Charness & Dave, 2017; Dave & Wolfe, 2003; but see Tentori et al., 2013).

As widespread as the confirmation bias is, possible explanations abound. Cognitive explanations, for example, assume that it is a result of limited cognitive resources leading to the application of heuristics or the inability to test alternative hypotheses in parallel (Nickerson, 1998). Motivational explanations, in contrast, assume that people prefer positive over negative thoughts and thus search for evidence that confirms their current beliefs (Dawson et al., 2002; Kunda, 1990). In the context of probability judgments, LaBella and Koehler (2004) concluded that averaging models alone cannot explain their findings. They proposed a modification that allows for subjective evaluation of the implication of evidence given previously encountered information. In a Bayesian context, inductive confirmation is used to explain certain aspects of the conjunction fallacy that are easily explained by averaging accounts but not by Bayesian models (Crupi et al., 2008; Tentori & Crupi, 2012). Inductive confirmation assumes a piece of evidence can change the credibility of a hypothesis and this change can be positive or negative (Tentori et al., 2013).

Order Effects

When judging a case, a jury may, for example, hear the prosecution's presentation before the defense. In such sequential-updating situations, Bayesian theory stipulates that the presentation order of the material should not influence the final judgments. In other words, whether the prosecution or the defense presents first ought not matter. In contrast, past work

has shown that people's judgments are sensitive to the presentation order of sequential evidence (see Hogarth & Einhorn, 1992, for an early overview and Trueblood & Busemeyer, 2011, for a more current review). This holds irrespective of the nature of the evidence, which could be marbles sampled out of an urn (Shanteau, 1970), risky monetary gambles (Hertwig et al., 2004), legal evidence in a jury trial (Furnham, 1986; Walker, Thibaut, & Andreoli, 1972), or clinical evidence in a medical case (Bergus et al., 1998). As an example, Bergus et al. (1998) investigated order effects in a medical scenario in which physicians had to judge the probability that a patient was suffering from a urinary tract infection (UTI). After an initial piece of information, this probability was judged to be approximately .60. Two more pieces of information followed: one that indicated a UTI and one that was inconclusive. If physicians received the indicative information last, their probability judgments for UTI were higher than if they received the inconclusive evidence last. Neither such recency effects, where people place more weight on the most recent piece of information, nor primacy effects (Hogarth & Einhorn, 1992), where people place more weight on the piece of information that they encountered first, are in line with Bayesian principles.

One possible explanation for order effects might be that people simply forget previous information or remember it incorrectly. Further, it has been found that rather than being memory based, closely spaced sequential judgments are made online with decision makers sequentially updating their judgments (Hastie & Park, 1986).

Theoretical explanations of order effects have thus focused on the sequential updating behavior that people seem to employ and that leads to these effects. Hogarth and Einhorn (1992) argued that order effects result from people updating their behavior according to a belief-adjustment model where new information is integrated with old information in a weighted-additive fashion. McKenzie et al. (2002) extended this model to account for increases and decreases of confidence in two sides of a dispute. This was done by adding a

piece of information's minimum acceptable strength as a reference point against which to measure the impact of the new piece of evidence.

Most recently, order effects have been explained by using a quantum theoretical approach (Busemeyer et al., 2011; White et al., 2013; Trueblood & Busemeyer, 2010). This model has been compared to Hogarth and Einhorn's (1992) adding-and-averaging belief-adjustment model and has explained people's behavior in a medical diagnostic and a jury decision-making task better than competitors. Trueblood & Busemeyer (2010) also criticized the adding-updating model, which is the better of the two models from 1992, for not being able to take relative strength of different pieces of evidence into account. Therefore, in our similarity-updating model we employ a weighted-additive updating process, which can be sensitive to the relative strengths of different pieces of evidence.

Other Psychological Findings Relevant for Probability Judgments

Converging Evidence

According to Bayes's theorem, probability estimates increase if evidence converges toward a hypothesis.[1] For example, if a physician is sequentially presented with converging evidence suggesting that a patient has pneumonia, their probability estimate that the patient has pneumonia should continually increase. However, according to psychological models that have been proposed, such as the belief-adjustment model (Hogarth & Einhorn, 1992), probability judgments can decrease in light of converging evidence. If the first piece of evidence points toward pneumonia with a probability of .8, for example, and the second does so with a probability of .6 (still pointing toward pneumonia), the resulting probability could be $.80 \times w + .60 \times (1 - w) = .74$ if the weight $w = .70$. This is lower than the initial probability of .80 and diverges from Bayesian principles. Because people's reactions to converging evidence allow for a quantitative distinction between updating models such as the belief-

adjustment model and Bayes's theorem, we also asked people to give probability judgments based on sequentially presented converging evidence.

Insensitivity to Sample Size

In probability theory and statistics the law of large numbers postulates that if an experiment is repeated a large number of times the average of the results will be close to the expected value. However, it has been shown that people are often insensitive to sample size and follow the so-called law of small numbers (Kahneman 2011; Tversky & Kahneman, 1971).

Cognitive models for probability judgments based on probability theory predict that with increasing sample size the variance in probability judgments should increase and the probability judgments should become more extreme (in favor of the correct hypothesis). In contrast to the similarity model, probability theory predicts that sample size should affect people's probability judgments. In Study 4 we manipulated sample size to rigorously test the models' predictions regarding sample sizes.

Regression to the Mean

A statistical principle related to the law of large numbers is regression to the mean (Galton, 1886). This phenomenon describes the situation when the first measurement of a random variable is extreme but subsequent measurements then fall closer to its expected value. Regression to the mean is present whenever there is an imperfect correlation between two measurements and can lead to the so-called narrative fallacy, where a logical link between two unrelated events is sought (Kahneman, 2011). In the context of the dilution effect this becomes important because a smaller second judgment could be closer to the mean than a first, potentially extreme judgment and could be explained by a regression to the mean effect. However, the dilution effect predicts that this decrease occurs only if a second probability judgment is nondiagnostic and not if it is diagnostic.

The Probability-Theory-Plus-Noise Model

The Bayesian model provides a normative solution for probability judgments, but it might not present a plausible cognitive model (Tversky & Kahneman, 1974; but see Chater et al., 2006; Griffiths et al., 2010; Sanborn, & Chater, 2016; Tenenbaum et al., 2006). One promising alternative model for probability judgments is the probability-theory-plus-noise (PT+N) model (Costello & Watts, 2014, 2016) that can explain a range of probability-judgment phenomena such as conservatism, subadditivity, the conjunction fallacy, and the disjunction fallacy. The PT+N model assumes that single probability judgments are the result of a sampling process that draws instances from memory. The probability of sampling an instance of a certain type A is the actual probability that an event of type A occurs, $p(A)$. However, there is a chance $d < .5$ that a sampled instance is read incorrectly, meaning that with probability d an event A is incorrectly read as $\neg A$.

Although the PT+N model can explain various effects, it cannot explain the dilution effect or the confirmation effect because failures to read an event A correctly are on average distributed symmetrically around the correct probability of this event occurring. Furthermore, the PT+N model assumes errors in memory processes, so it is unclear how it is applied to cases where memory is of less relevance. We used a modified PT+N model assuming that people also make perceptional errors so that presented information could be misperceived by the decision maker. Appendix A provides the details of the modified PT+N model.

Similarity as a Psychological Concept

Similarity is an important concept in psychological research and has been approached in various ways in a plethora of studies (e.g., Evers & Lakens, 2014; Goldstone & Son, 2005; Medin et al., 1993; von Helversen et al., 2014). Similarity helps people make sense of the world and understand relations in the world because things that are similar tend to behave

similarly (such as frogs and toads). Generally, similarities can be identified conceptually according to rules or more automatically on the basis of perceived relations between objects.

The four major approaches to similarity have long been geometric, feature-based, alignment-based, and transformational models (Goldstone & Son, 2005). In geometric models, similarities between items are represented in terms of a set of points organized in an *n*-dimensional metric space whereby the size of *n* is determined by the number of characteristics of the items in the space. According to feature-based models (e.g., Tversky, 1977), similarity is assessed by a feature-matching process in which common and distinctive feature items are weighted and added, whereas in alignment-based models items are compared not only by matching their features but also by determining how their features align with one another. Finally, transformational models assess similarities based on transformational distances, that is, the number of transformations needed to transform one item into the other. Recently, it has been speculated that instead of classic probability theory, quantum probability theory might explain similarity processes in cognition (Pothos & Busemeyer, 2013; Pothos et al., 2013; Pothos et al., 2015; Pothos & Trueblood, 2015; Trueblood et al., 2014). According to quantum probability models, probabilities are computed geometrically via the projection of state vectors onto different (cognitive) subspaces and by computing the squared length of such projections (Trueblood & Busemeyer, 2011).

Mechanisms for Belief Updating

When people have formed a judgment or belief and receive more evidence, how should they adapt their beliefs in light of new incoming information? According to Bayesian theory, belief updating starts with a prior belief, which is then updated on the basis of subsequent observations. For each observation the likelihood of observing this piece of evidence is computed. The final integration of this likelihood and the prior belief results in

the updated (posterior) belief. For probability updating, Bayesian theory prescribes updating a current belief with new information by integrating the new information and the prior judgment. With each new piece of evidence, this process repeats itself (for details see Appendix B).

According to Bayes's theorem, nondiagnostic information does not change a previous belief. If the second piece of evidence, for example, is just as likely under Hypothesis A as under Hypothesis B—that is, if the likelihood ratio of the second piece of evidence is 1—then the last term in the equation can be canceled out. According to the Bayesian model, nondiagnostic information does not change the previously computed probability. The Bayesian model also cannot account for order effects and predicts that converging evidence always leads to increasing probability judgments. Thus, the Bayesian model can predict neither the dilution effect nor the order effect. Further, previous tests of Bayesian models of human cognition have been criticized for evaluating the models only according to their ability to predict data instead of directly comparing them to other cognitive models (Bowers & Davis, 2012). Therefore, in the current work we provide such a comparative test.

An alternative approach for belief updating is based on quantum probability theory (Busemeyer et al., 2011; Trueblood & Busemeyer, 2011). According to the quantum probability models, belief updating is the result of a change in perspective modeled in a vector space. Each dimension in the vector space represents the joint probability of a hypothesis and a piece of evidence. A belief that a hypothesis is true is then a point in the vector space represented differently depending on one's perspective. A change in perspective is mathematically modeled with a unitary transformation that is not commutative and thus, leads to order effects.

Weighting-and-adding processes have been repeatedly demonstrated to describe people's controlled judgments well (Anderson, 1981, 1996; Juslin et al., 2008; Lopes, 1985,

1987; Roussel et al., 2002; Shanteau, 1970, 1972, 1975) and have recently been shown to describe conjunctive probability judgments well (Jenny et al., 2014; Nilsson et al., 2009, 2013). Even the integration of sensory input from different modalities is assumed to happen through a weighting-and-adding process (Ernst & Bülthoff, 2004). Hogarth and Einhorn (1992) have described belief updating with weighting-and-adding processes, which have also been discussed in the context of the dilution effect (LaBella & Koehler, 2004). Weighted-additive integration of new and previous information can also be formalized by reinforcement-learning principles. In reinforcement learning, it is assumed that people form expectancies or beliefs based on the information they have experienced in their environment. On the basis of these expectancies people predict new outcomes. Depending on the size of their prediction error, people then adjust their beliefs according to the new outcomes (Sutton & Barto, 1998). It has also been argued that human reward-directed learning follows reinforcement-learning principles at a neural level (Tobler et al., 2005). Reinforcement-learning and weighting-and-adding principles of belief updating do not (necessarily) conform to the information-integration principles prescribed by probability theory.

The Similarity-Updating Model

According to the proposed similarity-updating model,[2] people form probability judgments on the basis of similarity and updating processes, specifically, that the probability that a hypothesis is considered true is an increasing function of the similarity between the evidence and the hypothesis. When people are presented with new evidence, they update the previous probability judgment with a new probability judgment with an anchoring-and-adjustment process (Hogarth & Einhorn, 1992) that takes a weighted average of the two probabilities (informed by the configural weighted average model; see Juslin et al., 2008; Nilsson et al., 2013) and potentially adjusts the results based on a similarities bias (Hahn et al., 2010; von Helversen et al., 2014). According to this theory, the dilution effect occurs

because the updating process is governed by a weighting-and-adding process in which the two pieces of evidence are integrated by weighting and adding. Whenever the weight on the nondiagnostic second piece of information exceeds zero, the probability judgment will be diluted. In dilution tasks, people typically provide judgments after the presentation of each piece of information. In such tasks we could expect people to put more weight on the more recent (nondiagnostic) piece of information because it has been shown that step-by-step processing can produce recency effects (Hogarth & Einhorn, 1992)

Similarity Process

In a nutshell, the assumption is that people first form two similarity judgments, which are based on the distance between a piece of evidence and Hypothesis A and the distance between the same evidence and Hypothesis B and then compare these similarities to compute a probability judgment that Hypothesis A is true. According to our model, first a distance (e.g., Nosofsky & Johansen, 2000) is computed:

$$d_{ij} = \left[\sum_m w_m \times \left|x_{im} - x_{jm}\right|^r\right]^{1/r}, \quad (1)$$

where x_{im} is the value of the ith piece of evidence on a psychological dimension m, x_{jm} is the value of hypothesis j on the psychological dimension, w is the weight put on a certain dimension, and r defines the metric (r = 1 for the city block metric, r = 2 for the Euclidean metric). Our stimuli consisted of distributions of counts of qualitatively distinct objects (red, blue, and green cards in a card game). We did not find a reason to assume that these dimensions are weighted unequally. Therefore, as we have three different color dimensions w was fixed to 1/3 in our model, meaning that the dimensions were weighted equally. Similarity judgments based on such separable features are usually better described by a city block metric as opposed to a Euclidean metric (Shepard, 1987), which is why r was fixed to 1.

This distance is transformed into a similarity between the evidence and the hypothesis by a nonlinearly decreasing function:

$$s_{ij} = \exp(-c \times d_{ij}^{l}), \quad (2)$$

where c is a sensitivity parameter that determines the rate at which similarity declines with distance and p determines the form of the similarity gradient (l = 1 for the exponential similarity gradient, l = 2 for the Gaussian similarity gradient).[3] According to Shepard (1987), an exponential similarity gradient is preferred in the case of discriminable stimuli, which is why we fixed l to 1. Because no similar rationale is provided for fixing c, we estimated c on the basis of the data, allowing it to be ≥ 0.001. If c is high, then objectively small distances between evidence and hypotheses are judged as large, resulting in low similarity judgments. The similarity-updating model uses the general concept of similarity to model judgments; which of the two distance matrices (city block metric or Euclidean metric) and which form of the similarity gradient (exponential or Gaussian) is more appropriate depend on the types of stimuli used.

The similarities between the evidence and Hypothesis A (s_1) and the evidence and Hypothesis B (s_2) are then transformed into probabilities (Luce, 1959):

$$p(\mathrm{A}|E_1) = \frac{1}{1+e^{\theta(s_2-s_1)}}, \quad (3)$$

where E_1 is the first piece of evidence, s_1 and s_2 are the similarities between this evidence and Hypothesis A and Hypothesis B, respectively, and θ is a free parameter that determines how strongly Hypothesis A is favored over Hypothesis B if the evidence speaks for Hypothesis A. We allowed θ to range between 0 and 1,000. The larger this parameter's value, the more clearly Hypothesis A is favored—in other words, the more different the two similarities are judged to be. The smaller θ is, the more conservative (closer to .50) estimates of probabilities become.

Our model assumes that these initial probabilities are stored in a nonlazy way, meaning that they are abstracted from the samples and stored as probabilities (Lindskog et

al., 2013). At the time of presentation of the second sample, participants have access to a memory trace of the first probability.

Updating Process

In light of new information, two new similarity judgments are formed according to Equations 1 and 2 and transformed into probability judgments according to Equation 3. In the spirit of the belief-adjustment model proposed by Hogarth and Einhorn (1992), the two individual probability judgments are integrated by a cognitive model based on the principle of anchoring and adjustment (Chapman & Johnson, 2002; Epley & Gilovich, 2001, 2006; Tversky & Kahneman, 1974). Anchoring and adjustment has been used to explain human perspective taking (Epley et al., 2004), economic decisions (Joyce & Biddle, 1981; Wansink et al., 1998), and recently judgments from multiple cues (Albrecht et al., 2019), and risky choices (Millroth et al., 2019).

The anchor is modeled after the original weighting-and-adding process proposed by Hogarth and Einhorn (1992):

$$p_{\text{avg}}(\text{A}|E_1, E_2) = (1 - \tau) \times p(\text{A}|E_1) + \tau \times p(\text{A}|E_2), \quad (4)$$

where τ is a recency parameter with values between 0 and 1 that measures how much weight is put on the more recent piece of evidence, $p(\text{A}|E_2)$.[4] The τ parameter is comparable to the weight parameter in the value-updating model for risky choice (Hertwig et al., 2005).

The adjustment is based on the concepts of inductive confirmation (Carnap, 1962; Tentori et al., 2013) and follows the ideas sketched by LaBella and Koehler (2004). In a nutshell, if people believe Hypothesis A is likely to be true given evidence E_1 ($p(\text{A}|\text{E}_1) > 0.5$ according to Equation 3), they have a tendency to interpret a new piece of evidence E_2 relative to what they have learned about Hypothesis A through E_1. We assume that this confirmation mechanism is a result of the similarity-based processing underlying probability estimates (cf. Hahn et al. 2010; von Helversen et al., 2014). More precisely, if the new piece

of evidence is more similar to the favored hypothesis than the previous piece of evidence, that is, $(s_{E_2,\mathrm{A}} - s_{E_1,\mathrm{A}}) > 0$, then the hypothesis is *confirmed*, leading to an increase in the probability estimate. If, however, it is less similar and does not confirm the previous evidence, the probability estimate decreases. Mathematically, this similarity-based confirmation is modeled as

$$c(E_2, \mathrm{A}|E_1) = (s_{E_2,\mathrm{A}} - s_{E_1,\mathrm{A}}) * |(p(\mathrm{A}|E_2) - p(\mathrm{A}|E_1))|, \quad (5)$$

The similarity difference between the hypothesis and the two pieces of evidence is weighted by the absolute difference of the associated probabilities. This means that the larger the difference between the two probability judgments, the larger the impact of the similarity difference. The final predicted probability judgment is the averaged probability adjusted by the confirmation component:

$$p(\mathrm{A}|E_1, E_2) = \begin{cases} p_{\mathrm{avg}}(\mathrm{A}|E_1, E_2) + c(E_2, A|E_1) * (1 - p_{\mathrm{avg}}(\mathrm{A}|E_1, E_2)) & c(E_2, \mathrm{A}|E_1) \geq 0 \\ p_{\mathrm{avg}}(\mathrm{A}|E_1, E_2) + c(E_2, A|E_1) * p_{\mathrm{avg}}(\mathrm{A}|E_1, E_2) & \mathrm{c}(E_2, \mathrm{A}|E_1) < 0 \end{cases} \quad (6)$$

Figures C1 and C2 in Appendix C illustrate the anchoring-and-adjustment process.

Predictions

The similarity-updating model can account for a lot of findings from the probability judgment literature. The dilution effect is predicted if the second piece of evidence is nondiagnostic and the weight on this piece is larger than zero. Generally, the similarity-updating model judges evidence as nondiagnostic and produces a probability of .50 when at least one of the following three sufficient conditions hold: (a) when according to Equation 1, the distances between the evidence and hypotheses are identical; (b) when the sensitivity parameter *c* is large and produces identical similarities of 0 for all distances > 0; or (c) when θ in Equation 3 approximates 0, meaning that the difference between the two similarities is judged negligible.

The similarity-updating model also accounts for confirmation effects under certain circumstances. Specifically, if the second piece of evidence is nondiagnostic and has a higher perceived similarity to the preferred hypothesis than the first piece of evidence (assuming the difference in associated probabilities is not zero) the combined probability estimate can be higher than the first probability estimate, if the second piece of evidence is much more similar to the preferred hypothesis than the first piece of evidence (see Appendix C). This is due to the similarity-based confirmation mechanism and cannot be explained by standard averaging mechanisms (LaBella & Koehler, 2004).

The similarity-updating model is also sensitive to different types of nondiagnostic samples (LaBella & Koehler, 2004). If the nondiagnostic evidence is neutral, in the sense that it includes mainly information that is not represented in both hypotheses, the dilution effect is predicted to be higher (or more likely to happen) than if the nondiagnostic evidence is mixed, meaning that it includes information used to describe the hypotheses, where confirmation effects are more likely. The reason is that neutral nondiagnostic samples are on average not very similar to the hypotheses as they do not touch on the information presented there. Thus, the probability judgment is likely decreased by the similarity-based confirmation mechanism (Equation 5).

The similarity-updating model can also account for order effects such as recency and primacy, by the differential weighting of the different pieces of evidence in the model. Let us reconsider Bergus et al.'s (1998) medical scenario. Using a numerical example, it is easy to see how the similarity-updating model would predict the pattern observed in this scenario. Assume that the indicative information resulting from the similarity process indicates a UTI with a probability of .80 and the inconclusive information indicates a UTI with a probability of .50 and another illness with a probability of .50. If the indicative information is presented first, the intermediate probability is $.20 \times .60 + .80 \times .80 = .76$ (assuming $\beta = .80$) and the

final probability is .20 × .76 + .80 × .50 = .55. In contrast, if the indicative information is presented last, the intermediate probability is .20 × .60 + .80 × .50 = .52 (assuming β = .80) and the final probability is .20 × .52 + .80 × .80 = .74 and thus higher than before. This recency effect is predicted by the similarity-updating model. Further, in contrast to the Bayesian model and as discussed earlier, the similarity-updating model does not necessarily predict that judgments will continually increase in light of converging evidence.

A Judgment Paradigm

To test our similarity-updating model, we implemented the classic book-bags-and-poker-chips paradigm as a computerized card game (Edwards, 1968). We chose this paradigm with stochastic events to rule out that nondiagnostic evidence could carry semantic information about Hypothesis A or B (Peters & Rothbart, 2000). On each trial, participants received two novel card decks (A and B) and samples were drawn with replacement from a randomly chosen and undisclosed deck. Each deck totaled 100 cards and contained blue, red, and green cards. Above the cards, numbers between 10 and 80 indicated how many cards per color each deck contained Participants first received only one sample and were asked to judge which deck the sample had been drawn from by clicking one of two buttons on the keyboard. After their choice, they were additionally asked to estimate the probability that the chosen deck had generated the sample. Participants provided their probability estimate by moving a slider on a ruler marked with 50% (left end) and 100% (right end). To prevent anchor effects, the slider on the ruler appeared only upon clicking the ruler and disappeared again after participants indicated their probability of choice. A button below the ruler indicated the percentage the ruler was pointing to. Upon clicking this percentage, participants could proceed. Thereupon a second sample was drawn and presented below the first sample. Now, participants were asked to identify the deck that the two samples came from and state the probability that their deck was the source of the two samples.

As an example, assume that Deck A consists of 30% blue, 20% red, and 50% green cards and Deck B 20% blue, 30% red, and 50% green cards. Note that in both decks half of the cards are green cards. This means that a sample consisting of only green cards comes from either deck with the same likelihood and is therefore nondiagnostic. Further, as the percentages of red and blue cards are reversed in the two decks, a sample consisting of an equal number of red and blue cards would also be nondiagnostic. For instance, the first sample consists of two blue, one red, and four green cards and the second sample two blue, two red, and three green cards. Note that the second sample is nondiagnostic. All updating processes are assumed to remain the same with each additional sample. That is, the process of updating a first probability judgment with the information of a second sample is assumed not to differ from the process of updating a second probability judgment with the information of a third sample. Considering these decks and samples, what probability does the similarity-updating model assign to Deck A? According to the model, the distance between the first sample and Deck A is $.33\ [|.30–.29|^1 + |.20–.14|^1 + |.50–.57|^1] = .05$. This results in a similarity of $e^{-1 \times .05} = .95$ when $c = 1$. The distance between the second sample and Deck A is $.33\ [|.20–.29|^1 + |.30–.14|^1 + |.50–.57|^1] = .11$, resulting in a similarity of $e^{-1 \times .11} = .90$. Thus, the probability judgment of Deck A given the first sample is $1 / (1 + e^{6 \times (.90–.95)}) = .57$ for a value of $\theta = 6$. Following the same principle, the probability judgment for Deck A given the second sample is $1/ (1+e^{6 \times (.94–.94)}) = .50$. Updating the first probability judgment with the second one then results in a final probability judgment of $.40 \times .57 + .60 \times .50 = .53$ if the second piece of evidence were weighted with $\beta = .60$ and thus given more weight than the first piece of evidence. In contrast to the Bayesian model according to which the first estimate is not influenced by the second, nondiagnostic sample, the similarity-updating model decreases its initial probability estimate.

In Study 1, two samples of seven cards each were presented sequentially and both samples were visible on the screen during the whole trial. In Study 2, the first sample was removed from the screen when the second sample appeared. In Study 3, three samples were sequentially presented. In Study 4 half of the trials included samples with 14 instead of seven cards.

Model Comparison

To recap, the similarity-updating model makes different predictions from alternative models that we consider competitors: the Bayesian model and the PT+N model (Costello & Watts, 2014, 2016). We tested this in four different studies. First, the models make diverging predictions when it comes to converging evidence. Whereas the Bayesian model and the PT+N model predict that probabilities increase in light of sequentially presented converging evidence, the similarity-updating model predicts that probability judgments can decrease if later converging evidence is less predictive than earlier evidence. Second, the Bayesian model and the PT+N model predict on average no effects of presentation order of the combined judgments, because noise causes the probability judgments to be centered around the predicted mean. In contrast, the differential weighting of sequentially presented pieces of evidence leads the similarity-updating model to produce order effects. Third, the models predict different reactions to nondiagnostic information. The probability judgment does not change as a result of nondiagnostic information according to the Bayesian model and on average also does not change according to the PT+N model. The similarity-updating model predicts the dilution effect; that is, probabilities are diluted when nondiagnostic information is encountered. These diverging model predictions can be inferred directly from the models' structure. Fourth, the similarity-updating model is insensitive to changed sample size while the Bayesian model and the PT+N model predict judgments will become more extreme and have decreasing trial-by-trial variance with increasing sample size.

Quantitative Methods to Test the Similarity-Updating Model

In this article we have proposed the similarity-updating model, which combines two established single models: a similarity-based mechanism to obtain beliefs and a belief-updating mechanism based on averaging and adjustment. Our approach deviates in two respects from past research on probability judgments: We assume that (a) single probability judgments are the result of a similarity-based process instead of being conditional probabilities as specified in probability theory, and (b) belief updating takes place by an anchoring-and-adjustment process instead of a Bayesian-updating process. Appendix C shows the behavior of the adjustment process and Appendix D shows how a similarity mechanism differs from likelihoods specified by a Bayesian model.

To rigorously test our model, we quantitatively tested it against two competing models: a Bayesian model and the PT+N model. We additionally compared the models to a benchmark baseline model. Per person, this model takes the mean of the observed first probability judgments and always predicts this probability on all trials after one sample. And after two samples, it takes the mean of the observed second probability judgments. That is, if the mean of a person's probability judgments is 84%, for example, this model predicts on average 84% with the random error as specified above. Any cognitive model that claims to provide a good account of probability judgments needs to outcompete the baseline model as a plausibility check.

For all models we assumed an error process so that the predicted judgment would vary around the most likely point estimate (cf. Budescu et al., 1997; Juslin et al., 1997). We used a normalized truncated normal probability density likelihood function to link the models' point predictions with people's judgments. One of the implications of this error theory is that even if people generally follow the similarity-updating model, this does not

mean that they always show the dilution effect; their final judgment can also exceed or be equal to their first estimate.

In Studies 3 and 4 we additionally tested how generalizable the predictions of the different models are, by estimating the models on a subset of the data and generalizing the predictions of the models to the remaining data set. Study 3 tested the generalizability of the proposed belief-updating mechanism and Study 4 tested the generalizability across different sample sizes.

Study 1: Full Display of Samples

The dilution effect has been shown in many different domains (Hackenbrack, 1992; LaBella & Koehler, 2004; Macrae et al., 1992; McKenzie et al., 2002; Meyvis & Janiszewski, 2002; Peters & Rothbart, 2000; Shanteau, 1975; Shelton 1999; Smith et al., 1999; Troutman & Shanteau, 1977; Waller & Zimbelman, 2003). As a starting point, the goal of Study 1 was to replicate the dilution effect in our experimental setting and to contrast the similarity-updating model, the Bayesian model, and the PT+N model based on their qualitative predictions and in their predictive power relative to each other and to a baseline model.

Method

Participants

Twenty-five undergraduate students (Mdn_{age} = 22 years, 76% women, 24% men) at the University of Basel participated. Participants were compensated with either course credit or book vouchers worth 15 Swiss francs (CHF). Additionally, they received a performance-contingent bonus (*Mdn* = 2.10 CHF).

Materials

The experiments were computerized. Participants were presented with a diverse set of randomly ordered games involving two decks of cards. For the distributions of cards in the

first deck, all combinations of three underlying probabilities of 10%, 20%, 30%, 40%, 50%, 60%, 70%, and 80% were used. These probabilities—for example, 20% / 50% / 30% red, blue, and green cards—always added up to 100%. To construct the second deck, one of the probabilities was held constant and the other two switched positions, resulting in 30% / 50% / 20% red, blue, and green cards, for example.

The sample distributions for 81% of all trials were determined by randomly drawing seven times from a Dirichlet distribution with the underlying probability distribution of the picked deck. The Dirichlet distribution is a multivariate generalization of the beta distribution and takes a symmetrical (i.e., "uniform") shape when its parameters are all set to 1. Sampling values from a three-parameter Dirichlet distribution (with all three parameters set to 1) produces three values between 0 and 1 that sum up to 1. For 19% of all trials, one sample was randomly sampled and the other sample was tweaked such that its likelihood of being drawn from Deck A was identical to its likelihood of coming from Deck B. These nondiagnostic samples consisted of either only cards of the color that was equally represented in the two decks, or a certain number of this card and an equal number of the other two cards. To simplify the task for the participants the samples were sorted according to color. All games were presented twice, once with the original order and once with a switched order of samples. This means that the nondiagnostic samples were sometimes presented first, which allowed us to elicit single probability judgments for them. The first sample remained present at the time that the second sample appeared.

Procedure

Study 1 involved 86 rounds consisting of one game with two samples each. The two samples were presented sequentially, with both samples visible on the screen at the end of a trial. In each round, participants chose the deck that they thought was more likely to have generated the samples they drew. Additionally, they stated the probability that their chosen

deck and not the other one had generated the sample. At the end of the experiment, one round was randomly picked and participants won 2.50 CHF if they had chosen the right deck in that round. The participants received an additional reward for the probability judgment that they had provided after having seen two samples in that round. The reward was based on an inverse Brier score (Brier, 1950), which was calculated by subtracting the squared difference between the judged probability and the outcome score from 1. If, for example, the outcome score for Deck A was 1 (i.e., Deck A was the source of the samples) and the participant assigned this deck a probability of 70%, then the inverse Brier score was $1–(.70–1)^2 = .91$. This score was multiplied by 5 (e.g., = 4.55) and the resulting number was rounded to one decimal point and paid in CHF. Thus, in this example, the participant would have gained 2.50 CHF for the right choice and 4.60 CHF for the probability judgment. If the participant chose wrongly and assigned the probability to the wrong deck, the Brier score (e.g., $[.70–1]^2 = .09$) was multiplied by 5 CHF and the participant received 0 CHF for the incorrect choice and in this case 0.50 cents for the probability judgment.

After providing informed consent, participants read through the instructions on the screen. They additionally received a printed version of the instructions (see Appendix E), which they could hold onto throughout the experiment. In these instructions, the whole procedure was explained and it was stressed that the sampling always took place with replacement and that the distributions of decks were therefore not affected by the sampling procedure. Further, it was stressed that the probability judgment always concerned the chosen deck. Participants were finally informed that in the end, one of the trials they had played would be randomly picked and played out and that they could win an extra 2.50 CHF for their correct choice and, depending on the accuracy of their probability judgment, up to 5 CHF extra.

To familiarize themselves with the task, participants then played five trials, which did not yet count. After these training trials, participants were allowed to ask remaining questions concerning the task, if necessary. After everything was clear to the participants, the 86 rounds began. At the end, participants provided us with a check of whether they had understood the task by describing in writing how they had solved the task. Twenty-two of the 25 participants had understood the task. For three participants, it was not clear if they had understood that within one round both samples were drawn from the same deck.

Results

Participants' General Performance

After inspecting one sample, participants identified the correct deck in 77% of all trials. This performance improved to 83% correct choices after two samples were presented. In 20% of trials, participants changed their minds between the first and the second sample. The majority (65%) of these switches resulted in correct choices. Normatively, using likelihood and Bayes's theorem, the correct solution could be identified in 89% of all trials after the presentation of a first diagnostic sample, and participants identified it in 82% of all trials. The correct solution could not always be identified because the samples were noisy and sometimes happened to have a higher likelihood of coming from the deck that they were not drawn from. After the presentation of a second diagnostic sample, the normative account's success rate increased to 92% whereas our participants' success rate remained stable at 82%. The median Spearman correlation coefficient between participants' first probability judgment and the normatively correct probability was $\rho = .76$ and the root mean square deviation (RMSD) was 0.17. The correlation decreased to $\rho = .64$ and the RMSD increased to 0.21 after the second sample was presented. This probably resulted from the updating process that people used, which seems to have distorted the second probability judgments. Especially in dilution trials, the integration of nondiagnostic information distorted the second probability

judgments. Thus, although people's probability judgments differed considerably from the normative solution, they nevertheless correlated with it and allowed participants to choose the correct deck in most trials. After investigating participants' general performance we examined whether their behavior leading to a potential dilution effect would allow us to differentiate between the Bayesian model, the PT+N model, and the similarity-updating model on a qualitative level.

The Dilution Effect

We defined dilution trials as trials in which the second sample was nondiagnostic (its likelihood of coming from either deck was .50), participants' judgments after only the first samples were > .50 (treating it as diagnostic and leaving room for a decreased second judgment), and participants did not change their choice between samples. We focused on trials in which people did not change their choice, because choice changes based on nondiagnostic data when the first data were diagnostic could potentially be due to guessing and not only to perceiving a nondiagnostic piece of information as diagnostic or to a specific information-integration process.

In total, dilution trials accounted for 14% of all trials. The percentage of trials in which participants showed the dilution effect averaged around 64% (median; range: 8% to 100%). More than half (68%) of the participants showed the dilution effect in more than half of these trials. Sixty-three percent of these judgments were lower after participants saw the additional nondiagnostic sample than after they saw only a diagnostic sample. We performed a regression analysis to examine the dilution effect between the first and second probability judgment. If the participants ignored nondiagnostic information as prescribed by the Bayesian model and the PT+N, the slope of the regression line would have a value of 1, whereas a smaller value shows the dilution effect. In fact, the observed slope was .54, indicating a strong dilution effect. Figure 1 shows the difference between participants' first

and second probability judgments in all four studies, when the first sample, the second sample, or neither of the samples was nondiagnostic. When the second sample was diagnostic (because either both samples were diagnostic or the first sample was nondiagnostic) the majority of differences was positive, meaning that the second, combined probability judgment was greater than the first. When the second sample was nondiagnostic, however, the situation was reversed, with a clear majority of second probability judgments being smaller than the first, representing a dilution effect. This clearly shows that the dilution effect depends on the second sample being nondiagnostic and that the dilution effect cannot simply be explained as a regression to the mean effect.

There are different factors that have an impact on the dilution effect. When the first probability judgment was lower than or equal to the median probability judgment (74%) over all trials and participants, dilution effects were observed in 56% of the trials. When the first probability judgment was higher than the median, dilution effects were observed in 71% of the trials. Thus, the rate of dilution effects depended on the size of the first probability judgment.

We also found differences between neutral and mixed nondiagnostic samples (as defined by LaBella & Koehler, 2004). The percentage of trials in which participants correctly identified nondiagnostic samples if presented first by responding with a probability judgment of 50% averaged around 88% (median; range: 0% to 100%). How often participants detected the nondiagnostic sample as such varied with sample type. If the nondiagnostic sample consisted of only the type of card that was equally represented in both decks (neutral sample), participants correctly identified it in a median of 100% of the trials (interquartile range [IQR] = 90% to 100%). This performance was lower for nondiagnostic samples, which consisted of a mix of cards (mixed sample; *Mdn* = 67%, IQR = 17% to 83%). Thus, participants were well able to identify nondiagnostic samples as such if they were presented first. For neutral

nondiagnostic samples, dilution effects were observed in 64% of the trials, whereas for mixed nondiagnostic samples, dilution effects were observed in 61% of the trials. On average the dilution effect was also greater for neutral nondiagnostic samples (mean difference between the first and second probability judgment of -8%) compared to mixed nondiagnostic samples (mean difference of -3%) as predicted by the similarity-updating model. In line with LaBella and Koehler's studies (2004), we also found a median of 15% (IQR = 0% to 40%) of judgments exhibited confirmation effects. After mixed nondiagnostic evidence, we found a median of 40% (IQR = 0% to 50%) of judgments exhibited confirmation effects. Mixed nondiagnostic samples were on average more similar to both decks (and thus the correct deck) than neutral nondiagnostic samples. As a result, the similarity-based confirmation mechanism led to smaller increases or even decreases in probability judgments for neutral samples (Equation 5). This, together with the finding that people recognize nondiagnostic information correctly in the first sample, indicates that if participants also identified the second nondiagnostic decks as nondiagnostic, the integration of the two decks may have led to the dilution effect (LaBella & Koehler, 2004).

In sum, the results replicate the dilution effect and differences between different types of dilution trials. Additionally, our analyses clearly show that the dilution effect is not due to regression to the mean. The Bayesian model and the PT+N model cannot account for the dilution effect nor for different probability judgments following mixed and neutral nondiagnostic samples. Thus, the qualitative results support the similarity-updating model.

Averaging and Adjustment

Pure averaging models predict that all combined probabilities after the second sample is presented must lie between the single probability estimates for Samples 1 and 2. We could test this indirectly because each trial was presented twice to participants with a switched order of the samples. Thus, we could estimate the single probability assigned to a

configuration of cards presented as a second sample and compare it with a trial where this configuration was presented as a first sample. On average, 47% (median; IQR = 42% to 55%) of all combined judgments after participants saw two samples lay between the two single probability judgments, 11% (median; IQR = 8% to 21%) were smaller than both single probability judgments, and 36% (median; IQR = 30% to 45%) were larger. Thus, this finding is not completely in line with a pure averaging mechanism (cf., Equation 4).

The similarity-based confirmation mechanism predicts that judgments could be smaller or larger than the single probability estimates under the condition that a second sample is much more (or less) similar to the chosen deck than the first sample. To test this, we needed to include assumptions about the similarity between a deck and a sample, which depends on values of free parameters and usually differs between people and tasks. To produce similarity predictions we used the median parameter values we obtained by applying our similarity-updating model to the participants' data (Table 1). The correlation of the similarity-based confirmation mechanism's predictions and the difference between the two observed probability judgments within one trial is clearly positive ($r = .42$, $p < .01$). Figure 2 shows a graphical representation of this correlation across all four experiments. In sum, the similarity-based confirmation mechanism predicts the difference between first and second probability judgments well.

Order Effects

The Bayesian model and the PT+N model predict trial order invariance in that the order in which the samples are presented does not influence the final probability judgments. In contrast, the similarity-updating model predicts order effects, which result from the unequal weighting of the first and the second sample. Indeed, the median absolute deviation between the two second probability judgments was 16% (IQR = 13% to 17%) over all participants. Thus, the fact that participants' probability judgments were affected by sample

order speaks against the Bayesian model on a qualitative level (assuming probabilities are taken at face value).

Converging Evidence

We investigated converging evidence by looking at all trials in which both samples stemmed from one of the two decks with a likelihood of > .50, in which participants picked the same deck after both samples, and in which the Bayesian model picked the same deck after all samples. These trials constituted 53% of the whole set. In 83% (median) of the subset of trials (25% of all trials) in which according to the Bayesian solution the second sample was more diagnostic than the first one, participants' final probability judgments exceed their first judgments (range: 42% to 100%). In 36% (median) of the subset of trials (25% of all trials) in which the first sample was more diagnostic than the second one according to the Bayesian solution, participants' judgments decreased between the initial and the final estimate (range: 4% to 82%).

Thus, in contrast to the Bayesian solution but in line with the similarity-updating model, converging evidence did not always lead participants to increase their judgments, and how participants treated converging evidence depended on the presentation order of the samples. How participants treated converging evidence provided us with qualitative evidence in favor of the similarity-updating model.

Quantitative Model Comparison

We estimated all models on the basis of the participants' individual complete data using maximum likelihood estimation and compared models by the Bayesian information criterion (BIC; Schwarz, 1978) that takes model complexity into account. The BICs for the four competitor models are listed in Table 2 and the median optimal parameter values are listed in Table 1. The median value of the weight parameter τ was .57, indicating that the

second piece of information was weighted more than the first. In other words, we observed a recency effect.

The similarity-updating model clearly outperformed the competing models. The median BIC of the similarity-updating model across all participants was -237 while for the PT+N model it was -165. Also, the similarity-updating model explains the responses of 24 of the 25 participants best according to model selection based on BIC. Neither the Bayesian model (median BIC of -83) nor the baseline model (median BIC of 4) performed well in comparison. Thus, overall, the similarity-updating model described our data best. In sum, the first experiment provided strong evidence that people's probability judgments are better described with the similarity-updating model than with the PT+N or the Bayesian model.

Discussion

Study 1 shows qualitative and quantitative findings supporting the similarity-updating model as compared to the competing models. The dilution effect, effects of different types of nondiagnostic samples, the existence of order effects, and the results on converging evidence are all in line with the predictions of the similarity-updating model. A quantitative model comparison based on BIC supports these results.

In Study 1, the first sample was still present at the time of the presentation of the second sample. Thus, the updating process could be based on description and performed with perfect memory. However, this is not always the case as people often receive bits of information sequentially and have access to only one piece of evidence at a time and they have to remember previous information. To test our model also for such a task where information is presented in a truly sequential way, we conducted Study 2.

Study 2: Sequential Display of Two Samples

The goal of the second experiment was to replicate the results from Study 1 in a setting in which the two samples were never simultaneously presented. In classic dilution

studies (e.g., Troutman & Shanteau, 1977), the samples were drawn out of real boxes with beads and had to be replaced before drawing the new sample so that the proportions of beads in the boxes remained unchanged. Therefore in the second experiment, the setting was identical to that in Study 1 with the exception that the first sample disappeared upon the draw of the second sample. Thus, participants had to remember the first sample at the time of deciding which deck both samples were from and providing the accompanying probability judgment. This allowed us to investigate if we could replicate our findings in a truly sequential setting.

Method

Participants

Twenty-six undergraduate students (Mdn_{age} = 23.0 years, 58% women, 42% men) at the University of Basel participated and were compensated with either course credit or book vouchers worth 15 CHF. Additionally, they received a performance-contingent bonus (*Mdn* = 2.65 CHF).

Materials

In contrast to in Study 1, in Study 2, samples were presented in a truly sequential manner. Additionally, a screen between rounds announced the next round and instructed participants to start the next round by pressing the letter "w" (which stood for the German word *weiter*, which means "proceed" in this context).

Procedure

The procedure in Study 2 was identical to that in Study 1 except that the first sample disappeared when the second one was presented. All participants understood the task.

Results and Discussion

Participants' General Performance

After inspecting one sample, participants identified the correct deck in 77% of all trials. This performance improved to 83% correct choices after two samples were presented. In 17% of trials, participants changed their mind between the first and the second sample. The majority (69%) of these switches resulted in correct choices. Normatively, the correct deck could be identified in 89% of all trials after the presentation of a first diagnostic sample whereas the participants identified the correct deck in 82% of these trials. After the presentation of a second diagnostic sample, the success rate of the normative account increased to 92% whereas the participants' success rate remained stable at 84%. The median Spearman correlation coefficient between participants' first probability judgment and the correct probability according to the normative solution was $\rho = .73$ and the RMSD was 0.19. The correlation decreased to $\rho = .58$ and the RMSD increased to 0.22 after the second sample was presented. As in Study 1, although people's probability judgments differed considerably from the normative solution, they nevertheless correlated well with this solution and allowed participants to choose the correct deck in most trials.

The Dilution Effect

In Study 2, 14% of all trials were dilution trials. The percentage of dilution trials in which participants showed the dilution effect averaged around 67% (median; range: 7% to 93%). Of these judgments, 58% were lower after having seen the additional nondiagnostic sample than after having seen only a diagnostic sample. The slope of a simple linear regression was .52, illustrating the dilution effect. One participant showed a dilution effect in 0% of all trials, but this was because this person responded ".50" on all trials. More than half of the remaining participants (56%) showed the dilution effect in more than half of these trials.

As in Study 1, participants showed fewer dilution effects (49%) when their first probability judgment was lower than or equal to the median probability judgment over all

trials and participants ($Mdn_{\text{probability judgment}}$ = 74%) than when their first probability judgment was higher (68%). Further, participants showed dilution effects in 61% of the dilution trials if the nondiagnostic sample was neutral, but in 53% of all other dilution trials. The median downward adjustment of the probabilities after the presentation of the nondiagnostic piece of information in these trials was 11% (IQR = 5% to 23%). We could thus replicate the dilution effect irrespective of whether the initial samples were present when the subsequent samples were shown.

The median percentage of correctly identified nondiagnostic samples averaged around 81% (range: 6% to 100%). If the nondiagnostic sample was a neutral sample, participants correctly identified it in a median of 100% of the trials (IQR = 73% to 100%). This performance was lower for nondiagnostic mixed samples (*Mdn* = 50%, IQR = 33% to 100%). We found confirmation effects in a median of 11% of the trials (IQR = 0% to 30%) after neutral samples and a median of 33% (IQR = 20% to 60%) after mixed samples.

Averaging and Adjustment

On average, 44% (median; IQR = 36% to 55%) of all combined, second probability judgments lay between the two single, first probability judgments, as predicted by pure averaging models. Around 5% (median; IQR = 0% to 10%) were smaller than both single probability judgments and 46% (median; IQR = 31% to 58%) were larger.

The correlation of the similarity-based confirmation mechanism's predictions (based on median parameter values, see Table 1) and the difference between the two observed probability judgments within one trial again show a positive correlation (r = .39, p < .01; cf. Figure 2). Again, the similarity-based confirmation mechanism predicts the difference between first and second probability judgments well.

Order Effects

The median difference between the two second probability judgments was 17% (IQR = 14% to 19%) in Study 2 when the samples were presented in reversed order. Thus, in Studies 1 and 2, the fact that participants' probability judgments were affected by sample order speaks against the Bayesian model on a qualitative level.

Converging Evidence

In 54% of all trials both samples came from one of the two decks with a likelihood of > .50 according to the Bayesian solution; participants picked the same deck after both samples, and the Bayesian model picked the same deck after each sample. In the subset of these trials (25% of all trials) in which the second sample was more diagnostic than the first sample according to the Bayesian solution, participants' final probability judgments exceeded their first judgments 84% of the time (range: 0% to 100%). This proportion was similar to that in Study 1. When the first sample was the more diagnostic of the two (in 25% of all trials), participants' judgments decreased between the initial and the final estimate 26% of the time (median; range: 0% to 57%). Thus, in contrast to the Bayesian solution but in line with the similarity-updating model, converging evidence did not always lead participants to increase their judgments, and how participants treated converging evidence depended on the presentation order of the samples.

Quantitative Model Comparison

In Study 2, the first sample disappeared as soon as the second sample appeared. We hypothesized that this manipulation would not change the cognitive process behind people's updating behavior.

In Study 2 we estimated all models in the same ways as in Study 1. The median BIC over all participants for the four models was 5 for the baseline model, −84 for the Bayesian model, −145 for the PT+N model, and −240 for the similarity-updating model, indicating that the latter provided the best model fit. The median optimal parameter values of all models are

listed in Table 1. According to model selection based on the BIC, 25 of 26 participants were best described by the similarity-updating model and one person was best described by the Bayesian model. So the similarity-updating model not only is the best model overall but also describes the individual participants' first probability judgments best.

The value of the weight parameter τ was approximately .64, slightly higher compared to Study 1 (.57). The increase can be explained by the first sample not being visible during the second probability judgment. Again this indicates that the second piece of information was weighted more than the first and that we observed a recency effect, irrespective of whether the first sample was still present at the time of judgment.

Study 3: Sequential Display of Three Samples

Study 2 provided strong evidence that the similarity-updating process describes not only description-based updating well, but also truly sequential updating including a memory component, in that the first sample or the first judgment had to be retrieved from memory when making a final judgment. Additionally, we wanted to test whether the similarity-updating model could also describe a longer updating process well, one that is based on more than two samples. In Study 3, participants could update their probability judgments twice, as they received three samples in total. This allowed us to test the models against each other in a longer, more complex sequential kind of belief updating.

Method

Participants

Twenty-four undergraduate students (Mdn_{age} = 23.50 years, 79% women, 21% men) participated in Study 3. Participants were compensated with either course credit or book vouchers worth 15 CHF. Additionally, they received a performance-contingent bonus (*Mdn* = 2.30 CHF).

Materials

The experimental setup of Study 3 was identical to that of Study 2 with samples being presented sequentially. The only difference was that in each round, three samples were drawn and presented to the participants. As in Studies 1 and 2, participants were confronted with a diverse set of randomly ordered games containing all combinations of three underlying probabilities of 10%, 20%, 30%, 40%, 50%, 60%, 70%, and 80%. The sample frequencies for 75% of all trials were determined by drawing random samples of seven cards from a binary distribution with the underlying probability of the respective deck. For 25% of all trials, one sample was tweaked such that its likelihood of being drawn from Deck A was identical to the likelihood of being drawn from Deck B. A randomly determined quarter of the regular trials were repeated three times with different orders of the samples. All dilution trials, that is, trials consisting of one nondiagnostic sample, were repeated three times with the nondiagnostic sample appearing either first, second, or third.

Procedure

The procedure in Study 3 was identical to that of the previous studies and included 84 rounds plus five practice rounds. Samples were presented sequentially and a sample disappeared when a subsequent one was presented. Twenty-three of the 24 participants understood the task.

Results and Discussion

Participants' General Performance

In Study 3, after seeing one sample, participants identified the correct deck in 72% of all trials. This performance improved to 73% correct choices after two samples were presented and to 75% after three samples. In 29% of trials, participants changed their mind between the first and the second sample or between the second and the third. The majority (54%) of these switches resulted in correct choices. Normatively, the correct solution could be identified in 78% of all trials after the presentation of a first diagnostic deck whereas

participants identified the correct deck in 75% of these trials. After the presentation of a second and a third diagnostic sample, the normative account's success rate increased to 83% and 84%, respectively, whereas participants' success rate first remained stable at 74% and finally increased to 78%. The median Spearman correlation coefficient between participants' first probability judgment and the normatively correct probability was $\rho = .77$ and the RMSD was 0.19. The correlation decreased to $\rho = .60$ after the second sample was presented and to $\rho = .49$ after the third sample, whereas the RMSD decreased to 0.23 and 0.22, respectively. In sum, as in Studies 1 and 2, although people's probability judgments differed considerably from the normative solution, they nevertheless correlated well with this solution and allowed participants to choose the correct deck in most trials.

The Dilution Effect

In total, 22% of all trials were dilution trials, meaning that either the second sample was nondiagnostic, participants did not change their choice between Samples 1 and 2, and their first probability judgment was $\neq .50$; or the third sample was nondiagnostic, participants did not change their choice between Samples 2 and 3, and their second probability judgment was $\neq .50$. The percentage of trials in which participants showed the dilution effect averaged around 60% (median; range: 0% to 100%). The percentage of trials in which dilution effects were found was similar in the three studies (67%, 64%, and 60%). The slope of a simple linear regression between the first and the second probability judgment was .52 and the slope for a linear regression between the second and the third sample was .82, again illustrating the dilution effect.

In total, 67% of the participants showed the dilution effect in more than half of the dilution trials between the first (second) and the second (third) sample. The median downward adjustment of the probabilities after the presentation of a nondiagnostic piece of information in these trials was 7% (IQR = 3% to 14%). As in Studies 1 and 2, participants

showed fewer dilution effects (46%) when their first probability judgment was lower than or equal to the median probability judgment over all trials and participants (66%) than when their first probability judgment was higher (69%). Similarly, between Samples 2 and 3, they showed fewer dilution effects (45%) when their second probability judgment was lower than or equal to the median probability judgment over all trials and participants (68%) than when their second probability judgment was higher (55%). Further, if the nondiagnostic sample was neutral, dilution effects were observed in 60% of the dilution trials, whereas in all other dilution trials, dilution effects were observed between Samples 1 and 2 58% of the time, and 71% and 50% of the time between Samples 2 and 3, respectively. Thus, we could replicate the dilution effect not only irrespective of whether the initial samples were present when the subsequent samples were shown but also irrespective of the number of samples presented.

The percentage of correctly identified nondiagnostic samples averaged around 89% (median; range: 0% to 100%). If the nondiagnostic sample was neutral, participants correctly identified it in a median of 100% of the trials (IQR = 83% to 100%). This performance was lower for nondiagnostic mixed samples (*Mdn* = 84%, IQR = 54% to 100%). Similarly to the previous studies, we found confirmation effects in a median of 17% (IQR = 0% to 33%) of trials after neutral samples and a median of 31% (IQR = 11% to 51%) of trials after mixed samples in Study 3.

Averaging and Adjustment

In Study 3 not all trials were presented in reverse order, thus the following analysis is based on 44% of the data. On average, 46% (median; IQR = 25% to 53%) of all combined, second probability judgments lay between the two single, first probability judgments, as predicted by pure averaging models. Around 5% (median; IQR = 0% to 13%) were smaller than both single probability judgments and 48% (median; IQR = 37% to 62%) were larger.

The correlation of the similarity-based confirmation mechanism's predictions (based on median parameter values, see Table 1) and the difference between the two observed probability judgments within one trial again show a positive correlation ($r = .24$, $p < .01$; cf. Figure 2). Again, the similarity-based confirmation mechanism predicts the difference between first and second probability judgments well.

Order Effects

The median difference between the two second probability judgments was 7% (IQR = 6% to 9.00%) in Study 3. Thus, in all three studies, the fact that participant' probability judgments were affected by sample order speaks against the Bayesian model on a qualitative level.

Converging Evidence

In 14% of all trials in which all three samples were drawn from one of the two decks with a likelihood of > .50 according to the Bayesian solution, participants picked the same deck after all samples, and the Bayesian model picked the same deck after each sample. This low percentage of trials did not allow us to compute percentages for individual participants but only for overall participants. In the subset of trials (1% of all trials) in which according to the Bayesian solution the second sample was more diagnostic than the first one, and the third sample more diagnostic than the second, the percentage of trials in which participants' final probability judgments exceeded their first judgments and their second judgments was 81%. In the subset of trials (3% of all trials) in which the first sample was more diagnostic than the second and the second more diagnostic than the third according to the Bayesian solution, the percentage of trials in which participants' judgments decreased between the initial, the second, and the final estimate was 18%. Thus, in contrast to the Bayesian solution but in line with the similarity-updating model, converging evidence did not always lead participants to

increase their judgments, and how participants treated converging evidence depended on the presentation order of the samples.

Quantitative Model Comparison

Parameter estimation and model selection were done as in the previous two studies, including all probability judgments per trial. The median BIC over all participants for the four models was −29 for the baseline model, −47 for the Bayesian model, −245 for the PT+N model, and −406 for the similarity-updating model. The median best fit parameter values of all models are listed in Table 1. Twenty-one of 24 participants were best described by the similarity-updating model according to model selection based on the BIC, and three were best described by the PT+N model. So the similarity-updating model is the best model overall and also describes the individual participants' probability judgments best.

Study 3 also tested the predictive power of the similarity-updating model regarding generalizations to independent judgments discarded for parameter estimation. To this end, we estimated the model's parameters on the basis of the first two judgments in each trial and predicted the third judgment. The median deviance over all participants was 0 for the Bayesian model and −47 for the PT+N and for the similarity-updating model. According to model selection based on minimum deviance, the Bayesian model was chosen for three of 24 participants, the PT+N model for 10, and the similarity-updating model for 11 participants.

While the similarity-updating model and the PT+N model are about equally good in predicting participant's third probability judgment, the similarity-updating model is still the best model on an individual level and is also able to account for the qualitative patterns we found in all three studies.

Study 4: Varying Sample Size

Study 4's goal was to test whether sample size has an impact on probability judgments. The similarity-updating model, in contrast to the Bayesian or the PT+N model,

uniquely predicts that sample size should not influence the probability judgments because single similarity judgments are agnostic to sample size. Study 4's setting was identical to Study 2's. Around two thirds of the nondiagnostic and half of the diagnostic trials were chosen randomly and the samples were doubled to directly compare pairs of trials with 7 and 14 cards.

Method

Participants

Twenty-five students (Mdn_{age} = 25 years, 64% women, 36% men) at the University of Basel participated and were compensated with either course credit or 20 CHF. Additionally, they received a performance-contingent bonus with a maximum of 7.5 CHF.

Materials

Study 4 was based on the materials of Study 2. From Study 2 we randomly selected 60% of the nondiagnostic trials and 50% of the diagnostic trials. For each of the selected trials we tested one unchanged version (sample size of 7 cards) and one version with doubled sample size (14 cards). Around 5% of the decks with a sample size of 14 had to be changed because they inconsistently showed one or both decks with only 10 cards of a certain color but a sample with more than 10 cards with that color. In these tasks we switched two values in the decks so that the samples were consistent with the information given by the decks.

Procedure

The procedure in Study 4 was identical to that in Study 2. All participants understood the task.

Results and Discussion

Participants' General Performance

After inspecting one sample, participants identified the correct deck in 77% of all trials. This performance improved to 83% correct choices after two samples were presented.

In 20% of trials, participants changed their mind between the first and the second sample. The majority (66%) of these switches resulted in correct choices. Normatively, the correct deck could be identified in 89% of all trials after the presentation of a first diagnostic sample whereas the participants identified the correct deck in 82% of these trials. After the presentation of a second diagnostic sample, the normative account success rate increased to 92% whereas the participants' success rate remained stable at 83%. The median Spearman correlation coefficient between participants' first probability judgment and the correct probability according to the normative solution was $\rho = .73$ and the RMSD was 0.2. The correlation decreased to $\rho = .51$ and the RMSD increased to 0.25 after the second sample was presented. As in all the previous studies, although people's probability judgments differed considerably from the normative solution, they nevertheless correlated well with this solution and allowed participants to choose the correct deck in most trials.

Differences in Sample Size

All cognitive models that are based on probability theory and sampling in a broader sense predict differences between observations with different sample sizes. Bayesian models predict (independent of the assumed error component) a more extreme probability estimate for the correct deck. Frequentist models predict a lower variance. The similarity-updating model predicts no differences between sample sizes.

We first tested the predictions of the similarity-updating model (no difference due to sample size) against the predictions of frequentist models (lower variance) by using a Bayesian *t* test (Morey & Rouder, 2018). We found substantial evidence in favor of the null hypothesis, that there is no difference in the mean variance per participant between sample sizes, with a Bayes factor (BF) of $BF_0 = 3.55$ for the first probability judgments and $BF_0 = 3.56$ for the second probability judgments, as predicted by the similarity-updating model. The median standard deviation across participants for the first probability judgment was 12%

(IQR = 10% to 13%) for samples with seven cards and 12% (IQR = 9% to 14%) for samples with 14. For the second probability judgment, the standard deviation across participants was 10% (IQR = 8% to 13%) for samples with seven cards and 10% (IQR = 9% to 12%) for samples with 14.

Next, we tested the predictions of the similarity-updating model against the predictions of the Bayesian model, that there is a difference between the two sample sizes using a Bayesian *t* test on median judgments. The Bayesian *t* test gave substantial evidence in favor of the null hypothesis, that there is no difference between sample sizes. The BF_0 for the first probability judgment was 3.56 and for the second probability judgment also 3.6. These results support the similarity-updating model and speak against the Bayesian model and the PT+N model. The median first probability judgment across participants was 73% (IQR = 60% to 76%) for a sample size of 7 and 73% (IQR = 64% to 78%) for a sample size of 14. The median second probability judgment across participants was 75% (IQR = 63% to 82%) for a sample size of 7 and 75% (IQR = 64% to 85%) for a sample size of 14. All these results clearly show that there is no difference between the different sample sizes and speak in favor of the similarity-updating model and against the Bayesian model and the PT+N model.

The Dilution Effect

In Study 4, 16% of all trials were dilution trials. The percentage of dilution trials in which participants showed the dilution effect averaged around 75% (median; IQR: 47% to 94%). Of these judgments, 70% were lower after having seen the additional nondiagnostic sample than after having seen only a diagnostic sample. The slope of a simple linear regression was .72, illustrating the dilution effect. One of the participants showed a dilution effect in 0% of all trials, but this was because this person responded ".50" on all trials. More than two thirds of the participants (68%) showed the dilution effect in more than half of these trials. Further, if the nondiagnostic sample consisted of only the color that was equally

frequent in the two decks and was thus fairly obviously nondiagnostic, then participants showed dilution effects in 82% of trials, whereas in trials with mixed nondiagnostic second samples, dilution effects appeared in 75%.

The percentage of correctly identified nondiagnostic first samples averaged around 89% (median; IQR: 67% to 94%). If the nondiagnostic sample was a neutral sample, participants correctly identified it in a median of 100% of the trials (IQR = 75% to 100%). This performance was lower for nondiagnostic mixed samples (*Mdn* = 83%, IQR = 33% to 83%). We found confirmation effects in a median of 8% of the trials (IQR = 0% to 17%) after neutral samples and a median of 0% (IQR = 0% to 33%) after mixed samples.

Averaging and Adjustment

On average, 53% (median; IQR = 43% to 63%) of all combined, second probability judgments lay between the two single, first probability judgments, as predicted by pure averaging models. Only 11% (median; IQR = 4% to 18%) were smaller than both single probability judgments and 31% (median; IQR = 21% to 48%) were larger.

The correlation of the similarity-based confirmation mechanism's predictions (based on median parameter values, see Table 1) and the difference between the two observed probability judgments within one trial again show a positive correlation ($r = .23$, $p < .01$; cf. Figure 2). Again, the similarity-based confirmation mechanism predicts the difference between first and second probability judgments well.

Order Effects

The median difference between the two second probability judgments was 18% (IQR = 10% to 33%). Thus, as in Studies 1 and 2, the fact that participants' probability judgments were affected by sample order speaks against the Bayesian model on a qualitative level.

Converging Evidence

In 52% of all trials both samples came from one of the two decks with a likelihood of > .50 according to the Bayesian solution; participants picked the same deck after both samples, and the Bayesian model picked the same deck after each sample. In the subset of these trials (23% of all trials) in which the second sample was more diagnostic than the first sample according to the Bayesian solution, participants' final probability judgments exceeded their first judgments in 83% of these trials (median; range: 14% to 100%). This proportion was similar to that in Study 1. When the first sample was the more diagnostic of the two (in 25% of all trials), participants' judgments decreased between the initial and the final estimate in 32% of all trials (range: 0% to 96%).

Study 4 replicates the findings from Studies 1–3 that participants were affected by sample order and the type of nondiagnostic samples in dilution trials, and converging evidence did not always lead participants to increase their judgments. These findings hold even in a setting with partly increased sample sizes. Again, these results speak in favor of the similarity-updating model and against the Bayesian model and the PT+N model.

Quantitative Model Comparison

In Study 4 we estimated all models on the basis of participants' probability judgments following a maximum likelihood estimation approach. The median BIC over all participants for the four models was −3 for the baseline model, −55 for the Bayesian model, −121 for the PT+N model, and −262 for the similarity-updating model, indicating that the latter provided the best model fit. The median optimal parameter values of all models are listed in Table 1. According to model selection based on the BIC, 23 of 25 participants were best described by the similarity updating model, one was best described by the Bayesian model, and one by the PT+N model.

As a second test of the predictive power of the models we estimated parameters for all trials with a sample size of 7 and predicted the probability judgments for the trials with a

sample size of 14. The median deviances over all participants were −34 for the Bayesian model, −42 for the PT+N model, and −124 for the similarity-updating model. As in the model selection based on the BIC, 23 of 25 participants were best predicted by the similarity-updating model, one by the Bayesian model, and one by the PT+N model according to the minimum deviance.

To summarize the results of the four studies, we observed dilution effects in 60% or more of all dilution trials. When looking at all first and second probability judgments over all studies, the slope of a simple linear regression line was .58, illustrating the dilution effect. According to model selection based on the BIC, over 93% of all participants were best described by the similarity-updating model, 5% by the PT+N model, and 2% by the Bayesian model.

General Discussion

Research has shown that when people judge probabilities, they do not always behave according to Bayesian theory (Bar-Hillel, 1980; Edwards, 1968; Jenny et al., 2014; Nisbett et al., 1981; Tversky & Kahneman, 1983; but see Chater et al., 2006; Griffiths et al., 2010; Sanborn & Chater, 2016; Tenenbaum et al., 2006). Belief updating prescribes that beliefs and probability judgments should not be influenced by nondiagnostic information. In contrast, people's beliefs and judgments often change upon the presentation of nondiagnostic information. The dilution effect, a special case of this influence of nondiagnostic information on people's beliefs, has been observed in a plethora of studies in fields ranging from social reasoning to accounting. Few studies have provided thorough cognitive explanations as to why people show this behavior.

Introducing and Testing the Similarity-Updating Model

To explain the cognitive processes behind people's belief-updating behavior we developed a new cognitive model, the similarity-updating model, which is inspired by models

from the judgment and decision-making and categorization literature. The innovation of this model is its synthesis of cognitive models from two different fields of research: People's subjective probability judgments are modeled with similarity processes, which have been used in the categorization literature (Nosofsky & Johansen, 2000), and people's belief updating is modeled with weighting-and-adding processes, which have been used in judgment and decision-making research (Hogarth & Einhorn 1992; Jenny et al., 2014; Juslin et al., 2009; Nilsson et al., 2009, 2013). This model not only bridges different fields of research but also is based on concepts that have already been validated in the respective fields of research. Further, our model settles a debate started by Nisbett et al. (1981), in which they argued that similarity rather than averaging processes produce the dilution effect. In our view, it is the fact that people seem to follow a combination of both similarity and averaging processes when making their judgments that leads them to produce the dilution effect. The subtleties of this only become clear when the two processes are combined into one overarching model. The combination shows that although similarity processes can explain the individual judgments, the updating process of weighting and averaging eventually leads to the dilution effect. Our model also accounts for the criticism by Trueblood and Busemeyer (2011) that adding and averaging models cannot account for the strength of different pieces of evidence relative to each other, which our model does with the addition of the similarity-based confirmation mechanism.

The original belief-adjustment model uses a first probability judgment as the anchor and adjusts it relative to the new evidence (Hogarth & Einhorn, 1992). Our similarity-updating model deviates from this as it assumes that the anchor is the average of two probability judgments and the adjustment is actually a similarity bias (Hahn et al. 2010; von Helversen et al., 2014). This new formulation of the original belief-updating model helps us explain why people sometimes show the dilution effect but sometimes show a confirmation

effect instead. It depends on how similar the different pieces of evidence are relative to a preferred hypothesis.

We tested this model thoroughly within a cognitive modeling framework by estimating it on the basis of people's probability judgments. We tested the model on a trial-by-trial level for each individual participant. An additional strength of our model test is that we compared the model's ability to predict people's behavior to a Bayesian model, the probability-theory-plus-noise (PT+N) model, and a random baseline model. The models were tested against each other in four experimental studies, in which people provided subjective probability judgments and revised them in light of additional information in a card game. The task was to assess from which of two decks two or three sequentially presented samples of cards originated.

Although people's probability judgments differed considerably from the normative solution, they nevertheless correlated well with this solution and they were able to choose the correct deck in most trials. Consistently over all studies and participants, we observed dilution trials at a rate of approximately 60%. The greater participants' first probability judgments, the more dilution effects they produced. Overall, participants correctly identified a nondiagnostic first deck as such. Extrapolating from this and assuming that they also were quite well able to identify a second nondiagnostic deck, it seems likely that it was people's belief-updating mechanism rather than the way they judged individual probabilities that led them to produce the dilution effect. On average, their second probability judgment was also smaller only when the second sample was nondiagnostic. This suggests that this not a falsely qualified regression to the mean effect, which would be explained by the PT+N model.

The process behind people's first and subsequent probability judgments in this task was best described by the similarity-updating model as compared to a Bayesian model, the PT+N model, and a random baseline model. The similarity-updating model not only provided

the best model fit over all participants but also described the individual behavior of almost all participants best. These results held irrespective of whether the first sample was still present or removed at the time of the second sample being presented, for updating situations in which participants could make use of an additional third sample, and also for different sample sizes. A generalization test in Studies 3 and 4 additionally showed that the superior model fit was not due to overfitting or model flexibility that was too great.

One might think that a model that assumes diagnostic and nondiagnostic information is weighted and averaged and can predict the dilution effect would also always lead to the dilution effect. In contrast, there are several factors that led participants whose behavior followed the similarity-updating model to show the dilution effect at a rate of < 1. One reason is that the nondiagnostic piece of information is not always judged to be nondiagnostic by the similarity-updating model. In other words, when the second sample is falsely seen as diagnostic (instead of nondiagnostic), people show the dilution effect but think that they have used "diagnostic" information appropriately. A second factor is a post hoc increase in the probability judgments based on the averaging process due to a similarity-based confirmation mechanism. If the second piece of evidence is very similar to the preferred hypothesis, people might not show a dilution effect but rather a confirmation effect even if the evidence is nondiagnostic. Another factor is that people's behavior is not always consistent with the model but fluctuates around the model's predictions. In short, dilution effect rates of < 1 can arise due to the similarity-updating model not always judging nondiagnostic evidence as such and due to people not adhering to the model perfectly.

The similarity-based confirmation mechanism provides a possible explanation for the results obtained by LaBella and Koehler (2004) that people show no dilution effect if nondiagnostic evidence is neutral and even a confirmation effect if nondiagnostic evidence is mixed in a belief-updating task. We argue that this finding is a result of the overall similarity

between the second, nondiagnostic piece of evidence and a preferred hypothesis. In our experiment, we found a dilution effect for both mixed and neutral samples but it was smaller for mixed samples. We explain this by the fact that in our experiments, nondiagnostic samples were on average not very similar to either hypothesis and, thus, would in most cases not lead to a confirmation effect.

The similarity-updating model also makes the unique prediction that the sample size should not affect people's judgments. The Bayesian model and the PT+N model predict that probability judgments become more extreme and less noisy with increasing sample size. The results of Study 4 confirm the prediction of the similarity-updating model. The idea that in probability judgments people give too little weight to sample sizes has a long tradition in psychology (Tversky & Kahneman, 1971) and has also been observed in more recent work (e.g., Hoffart, Rieskamp, & Dutilh, 2019; Hoffart, Olschewski, & Rieskamp, 2019).

Characteristic Effects in Belief Updating

Above, we discussed that we found (a) dilution effects, (b) confirmation effects, (c) effects of diagnosticity, (d) order effects, (e) effects of converging evidence, and (f) insensitivity to sample size in our data and that these effects allowed us to differentiate between the similarity-updating model, the Bayesian model, and the PT+N model on a quantitative level.

Implications

Subjective Probability Judgment

Among other areas, similarity has previously informed research on quantitative estimations (Juslin et al., 2008; von Helversen & Rieskamp, 2009; von Helversen et al., 2014). In line with related findings (Nilsson et al., 2005; Read & Grushka-Cockayne, 2011), we have shown that introducing the concept of similarity to the study of subjective probability judgments provides important insights. With the similarity-updating model, we

tested a specific instantiation of similarity and contrasted it to other instantiations such as the similarity heuristic (Read & Grushka-Cockayne, 2011), representativeness as prototype similarity (Nilsson et al., 2005), and representativeness as relative likelihood (Nilsson et al., 2005), as well as alternative versions of the similarity-updating model. To reduce the length of this paper we have not reported the model comparisons but it turned out that the similarity-updating model was superior to these variants. Alternative models of similarity are the evidential support accumulation model (Koehler et al., 2003), cue-based relative frequency, and the probabilities from exemplars (PROBEX) model (Juslin & Persson, 2002). The first two models were not tested in the present studies because there is not a straightforward application of these models to our task. Further, applied to our task, which does not involve memory processes across trials, the PROBEX model boils down to the probability judgment part of the similarity-updating model with a probability judgment function, which differs slightly from Equation 3. This makes testing this model superfluous as long as no memory processes are involved.

Belief Updating

Weighting-and-adding processes have been successfully applied to explain people's behavior in conjunctive probability estimation (Jenny et al., 2014; Nilsson et al., 2009, 2013) and they have been demonstrated to perform better than normative as well as alternative cognitive models. In the present article we have shown that they can also well describe people's belief-updating processes, which have previously been described with belief-updating models (Hogarth & Einhorn, 1992) and the sigma model (Juslin et al., 2008). Our analyses show that the similarity-updating model, with its weighting and adding processes, outperforms alternative updating models.

A potential alternative approach to modeling belief updating is quantum probability theory (QPM; Busemeyer et al., 2011; Trueblood et al., 2017) and there is evidence

suggesting that the QPM explains several nonnormative behaviors. The QPM can be seen as a generalized and relaxed version of Bayesian probability theory that explains nonnormative behaviors usually with a context or background that gives rise to certain mathematical representations. For example, order effects are explained by the idea that people adopt different perspectives when evidence is presented and a shift between contexts is a noncommutative operation (Trueblood & Busemeyer, 2011). However, the dilution effect, as usually tested with the book-bags-and-poker-chips task, does not include different contexts or backstories. As such, theorizing a QPM that explains the dilution effect is not straightforward.

Similarity-Based Confirmation

Confirmation effects have a long-standing tradition in psychology and economics and have been shown to affect people's judgments and decision making in a variety of contexts (Jones & Sugden, 2001; Lord et al., 1979; Plous, 1991; Wason, 1968). However, computational cognitive models that intend to explain this effect are rare. Similarly, confirmation effects have rarely been investigated in the domain of probability judgments (but see LaBella & Koehler, 2004). We propose a similarity-based confirmation mechanism that is a part of the belief-updating process. Following an anchoring-and-adjustment process (Chapman & Johnson, 2002; Epley & Gilovich, 2001, 2006; Hogarth & Einhorn, 1992; Tversky & Kahneman, 1974), people average the probabilities of two sequentially presented pieces of evidence (Hogarth & Einhorn, 1992) but then adjust the probability as a result of a similarity bias (Hahn et al., 2010; von Helversen et al., 2014). The more similar the second piece of evidence is to a favored hypothesis, the higher the adjustment and, thus, the resulting probability judgment. This mechanism is very similar to the original belief-adjustment model but assumes that the anchor is given by the combination of the two independent probability values and not by the first probability. Thus, it can be viewed as an extension to the averaging

mechanisms, which are able to explain the dilution effect and other effects from the literature on probability judgments, to also account for confirmation effects (defined as the reverse of the dilution effect), thereby providing a theoretical idea of why and how dilution and confirmation effects interact in the domain of probability judgments.

Reduced Dilution by Expertise

It has been shown that the dilution effect decreases with expertise in several domains (e.g., Shelton, 1999; Smith et al., 1999). The similarity-updating model linked to some ideas from the distortion of information literature is also able to explain the absence of dilution, that is, no decrease in probability judgments in the face of nondiagnostic information. According to the similarity-updating model, the magnitude of decrease after a second sample has been presented depends on the perceived importance of the second sample. If this importance is zero, the probability judgment does not change when the second sample is presented. A decrease in the perceived importance thus implies a reduction in the observed dilution in probability judgments. For example, an experienced auditor or a judge might consider the importance of a new piece of evidence as rather small in comparison to a body of evidence already gathered.

Bridging the Gap Between Related Fields

To what extent can the similarity-updating model be applied to other domains? In general, it seems that the model can be applied whenever the probability for a certain event has to be judged given multiple pieces of information. One such domain is causal reasoning (Trueblood & Pothos, 2014; Trueblood et al., 2017). After learning a causal structure (aspect *A*/*B* causes event *E*) participants were asked to judge how likely event *E* is to occur given the presence/absence of aspects *A* and *B*. The similarity-updating model can also account for some of the effects in causal reasoning. Order effects in causal reasoning can be explained by the recency component in the updating mechanism. Similarly, the memoryless effect (the

probability that an aspect is present depends on only the most recent information; Trueblood & Pothos, 2014) can be explained by the recency component as well.

Another related domain is the literature on probability judgments of conjunctive events (e.g., Tversky & Kahneman, 1983). Here participants are asked to combine p(A), that is, the probability of A, and p(B) into a judgment of p(A and B) and the typical finding is that p(A and B) tends to fall between the two constituent probabilities, that is, if p(A) < p(B), then the typical finding is that p(A) < p(A and B) < p(B). This judgment pattern is typically referred to as the conjunction error, or the conjunction fallacy, as it violates the conjunction rule of probability theory. Notably, the similarity-updating model can predict the conjunction error with its averaging mechanism (for similar arguments see, Fantino et al., 1997; Nilsson et al. 2009, 2013).

Bridging Cognitive Psychology and Judgment and Decision-Making Research

Applying the concept of similarity to the area of judgment and decision making via the similarity-updating model fits the movement of applying concepts from more basal processes such as perception to higher order processes such as similarity judgments. This so-called mindful judgment and decision-making research has led to a more detailed understanding of judgment and decision-making phenomena (Johnson & Weber, 2009), for example, by modeling recognition within the cognitive architecture ACT-R (Schooler & Hertwig, 2005), modeling confidence judgments with evidence-accumulation models (Pleskac & Busemeyer, 2010), or modeling forced-choice decision making with evidence-accumulation models (Lee & Cummins, 2004).

Considering the similarity-updating model in the context of the dilution effect facilitated the realization that items that a researcher or experimenter might use because they are nondiagnostic will not necessarily be perceived and treated as nondiagnostic by the experiment participants. Although in our experiment nondiagnostic first samples were often

perceived as such, nondiagnostic second samples still influenced the initial probability judgments based on first diagnostic samples. This was due to people's tendency to put considerable weight on the second piece of (potentially nondiagnostic) evidence. Thus, by looking at the first probability judgments based on nondiagnostic samples and by inspecting the weight parameter of the estimated similarity-updating model (β), it is possible to distinguish whether the dilution effect is caused by distorted probability estimation or by the belief-updating mechanism. The dilution effects should always occur when the similarity-updating model identifies the nondiagnostic piece of evidence as such and weights it with a weight > 0.

Open Questions

Belief updating can occur in an immediate, online fashion where one's belief changes within seconds. Alternatively, it can also occur much more slowly, over the course of hours, days, or longer periods of time. The structure of the typical dilution task allows one to consider belief updating formed in a rapid sequence. Updating a belief in such an online fashion is an important skill. Imagine a security guard screening potential troublemakers at an airport. The security guard might first notice that a potential subject looks around quite nervously and therefore pays more attention to this person. Next, the guard sees a small child who is yelling for her dad and realizes that the potential subject is probably just looking for his child. In an alternative scenario, if the guard noticed that the nervous man was making secret signs to another person, the guard might become more suspicious about this person. In situations such as this it is crucial that people can accurately and quickly update their beliefs based on sequentially observed information and make good corresponding decisions. Thus, understanding how people form such belief updates crucially broadens the understanding of complex human cognition as a whole. This raises the question of if this model also explains

people's belief updating well if the updating happens over a longer period of time. This is an important question to address in future research.

Conclusions

People's probability judgments in a belief-revision task in which they experience the occurrence rate of events through sampling can be better described by a similarity-updating model consisting of similarity and weighting-and-adding processes than by a Bayesian model or alternative accounts of similarity (Nilsson et al., 2005; Read & Grushka-Cockayne, 2011) and updating processes (Hogarth & Einhorn, 1992; Juslin et al., 2008). It seems that people show the dilution effect because when forming and updating their probability judgments, they use a rule that lets them integrate nondiagnostic information into the judgment via a weighting-and-adding process. These findings are in line with previous findings (e.g., Anderson, 1981; Hogarth & Einhorn, 1992; Nilsson et al., 2009; Shanteau, 1970) that similarity and weighting-and-adding processes affect people's probability judgments. Although following a similarity-updating process leads people to take nondiagnostic information into account and produce the dilution effect, it still leads them to make generally good predictions and receive good decision outcomes, especially with decreasing noise. The similarity-updating model can describe the underlying cognitive process of people's probability judgments that often lead to accurate decisions despite violating probability theory.

References

Albrecht, R., Hoffmann, J. A., Pleskac, T. J., Rieskamp, J., & von Helversen, B. (2019). Competitive retrieval strategy causes multimodal response distributions in multiple-cue judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(6), 1064–1090. https://doi.org/10.1037/xlm0000772

Anderson, N. H. (1981). *Foundations of information integration theory*. Academic Press.

Anderson, N. H. (1996). *A functional theory of cognition.* Erlbaum.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*, 211–233. https://doi.org/10.1016/0001-6918(80)90046-3

Bergus, G. R., Chapman, G. B., Levy, B. T., Ely, J. W., & Oppliger, R. A. (1998). Clinical diagnosis and order of information. *Medical Decision Making, 18*, 412–417. https://doi.org/10.1177/0272989X9801800409

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*, 389–414. https://doi.org/10.1037/a0026450

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*, 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Budescu, D. V., Erev, I., & Wallsten, T. S. (1997). On the importance of random error in the study of probability judgment. Part I: New theoretical developments. *Journal of Behavioral Decision Making*, *10*, 157–171. https://doi.org/10.1002/(SICI)1099-0771(199709)10:3<157::AID-BDM260>3.0.CO;2

Busemeyer, J. R., Pothos, E. M., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychological Review*, *118*, 193–218. https://doi.org/10.1037/a0022542

Carnap, R. (1962). *Logical foundations of probability* (2nd ed.). University of Chicago Press.

Chapman, G. B., & Johnson, E. J. (2002). Incorporating the irrelevant: Anchors in judgments of belief and value. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 120–138). Cambridge University Press. https://doi.org/10.1017/CBO9780511808098.008

Charness, G., & Dave, C. (2017). Confirmation bias with motivated beliefs. *Games and Economic Behavior*, *104*, 1–23. https://doi.org/10.1016/j.geb.2017.02.015

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations [Editorial]. *Trends in Cognitive Sciences, 10*(7), 287–291. https://doi.org/10.1016/j.tics.2006.05.007

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review, 121*(3), 463–480. https://doi.org/10.1037/a0037010

Costello, F., & Watts, P. (2016). People's conditional probability judgments follow probability theory (plus noise). *Cognitive Psychology*, *89*, 106–133. https://doi.org/10.1016/j.cogpsych.2016.06.006

Crupi, V., Fitelson, B., & Tentori, K. (2008). Probability, confirmation and the conjunction fallacy. *Thinking & Reasoning*, *14*, 182–199. https://doi.org/10.1080/13546780701643406

Dave, C., & Wolfe, K. W. (2003). On confirmation bias and deviations from Bayesian updating. https://www.researchgate.net/profile/John_Duffy3/publication/229013128_On_confirmation_bias_and_deviations_from_Bayesian_updating/links/0046351c994d3dc7cb000000.pdf

Dawson, E., Gilovich, T., & Regan, D. T. (2002). Motivated reasoning and performance on the Wason Selection Task. *Personality and Social Psychology Bulletin, 28*(10), 1379–1387. https://doi.org/10.1177/014616702236869

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review, 106*, 180–209. https://doi.org/10.1037/0033-295X.106.1.180

Dougherty, M. R. P., & Hunter, J. (2003). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition, 31*, 968–982. https://doi.org/10.3758/BF03196449

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17–52). Wiley.

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science*, *12*(5), 391–396. https://doi.org/10.1111/1467-9280.00372

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science*, *17*(4), 311–318. https://doi.org/10.1111/j.1467-9280.2006.01704.x

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*(3), 327–339. https://doi.org/10.1037/0022-3514.87.3.327

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences, 8*, 162–169. https://doi.org/10.1016/j.tics.2004.02.002

Evers, E. R. K., & Lakens, D. (2014). Revisiting Tversky's diagnosticity principle. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.00875

Fantino, E., Kulik, J., Stolarz-Fantino, S., & Wright, W. (1997). The conjunction fallacy: A test of the averaging hypotheses. *Psychonomic Bulletin and Review, 4*, 96–101.

Furnham, A. (1986). The robustness of the recency effect: Studies using legal evidence. *Journal of General Psychology, 113*, 351–357. https://doi.org/10.1080/00221309.1986.9711045

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland, 15*, 246–263.

Goldstone, R. L., & Son, J. Y. (2005). Similarity. In K. J. Holyoak & R. G. Morrison (Eds.), *Handbook of thinking and reasoning* (pp. 13–36). Cambridge University Press.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences, 14*(8), 357–364. https://doi.org/10.1016/j.tics.2010.05.004

Hackenbrack, K. (1992). Implications of seemingly irrelevant evidence in audit judgment. *Journal of Accounting Research, 30*, 126–136.

Hahn, U., Prat-Sala, M., Pothos, E. M., & Brumby, D. P. (2010). Exemplar similarity and rule application. *Cognition, 114*(1), 1–18. https://doi.org/10.1016/j.cognition.2009.08.011

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review, 93*, 258–268. https://doi.org/10.1037/0033-295X.93.3.258

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*, 534–539. https://doi.org/10.1111/j.0956-7976.2004.00715.x

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2005). The role of information sampling in risky choice. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 72–91). Cambridge University Press.

Hoffart, J. C., Olschewski, S., & Rieskamp, J. (2019). Reaching for the star ratings: A Bayesian-inspired account of how people use consumer ratings. *Journal of Economic Psychology*, *72*, 99–116. https://doi.org/10.1016/j.joep.2019.02.008

Hoffart, J. C., Rieskamp, J., & Dutilh, G. (2019). How environmental regularities affect people's information search in probability judgments from experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(2), 219–231. https://doi.org/10.1037/xlm0000572

Hogarth, R., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*, 1–55. https://doi.org/10.1016/0010-0285(92)90002-J

Hotaling, J. M., Cohen, A. L., Shiffrin, R. M., & Busemeyer, J. R. (2015). The dilution effect and information integration in perceptual decision making. *PLoS ONE, 10*(9), Article e0138481. https://doi.org/10.1371/journal.pone.0138481

Jenny, M. A., Rieskamp, J., & Nilsson, H. (2014). Inferring conjunctive probabilities from noisy samples: Evidence for the configural weighted average model. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *40*, 203–217. https://doi.org/10.1037/a0034261

Johnson, E. J., & Weber, E. U. (2009). Mindful judgment and decision making. *Annual Review of Psychology*, *60*, 53–85. https://doi.org/10.1146/annurev.psych.60.110707.163633

Jones, M., & Sugden, R. (2001). Positive confirmation bias in the acquisition of information. *Theory and Decision*, *50*(1), 59–99.

Joyce, E. J., & Biddle, G. C. (1981). Anchoring and adjustment in probabilistic inference in auditing. *Journal of Accounting Research, 19*(1), 120–145.

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition, 106,* 259–298. https://doi.org/10.1016/j.cognition.2007.02.003

Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review, 116*, 856–874. https://doi.org/10.1037/a0016979

Juslin, P., Olsson, H., & Björkman, M. (1997). Brunswikian and Thurstonian origins of bias in probability assessment: On the interpretation of stochastic components of judgment. *Journal of Behavioral Decision Making*, *10*, 189–209. https://doi.org/10.1002/(SICI)1099-0771(199709)10:3<189::AID-BDM258>3.0.CO;2-4

Juslin, P., & Persson, M. (2002). PROBabilities from Exemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science, 26,* 563–607. http://dx.doi.org/10.1016/S0364-0213(02)00083-6

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Koehler, D. J., White, C. M., & Grondin, R. (2003). An evidential support accumulation model of subjective probability. *Cognitive Psychology, 46,* 152–197. https://doi.org/10.1016/S0010-0285(02)00515-7

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

LaBella, C., & Koehler, D. J. (2004). Dilution and confirmation of probability judgments based on nondiagnostic evidence. *Memory & Cognition*, *32*, 1076–1089. https://doi.org/10.3758/BF03196883

Lee, M. D., & Cummins, T. D. R. (2004). Evidence accumulation in decision making: Unifying the "take the best" and the "rational" models. *Psychonomic Bulletin & Review, 11*, 343–352. https://doi.org/10.3758/BF03196581

Lindskog, M., Winman, A., & Juslin, P. (2013). Calculate or wait: Is man an eager or a lazy intuitive statistician? *Journal of Cognitive Psychology*, *25*, 994–1014. https://doi.org/10.1080/20445911.2013.841170

Lopes, L. L. (1985). Averaging rules and adjustment processes in Bayesian inference. *Bulletin of the Psychonomic Society*, *23*, 509–512. http://dx.doi.org/10.3758/BF03329868

Lopes, L. L. (1987). Procedural debiasing. *Acta Psychologica*, *64*, 167–185. https://doi.org/10.1016/0001-6918(87)90005-9

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098–2109. https://doi.org/10.1037/0022-3514.37.11.2098

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* Wiley.

Macchi, L., Osherson, D., & Krantz, D. H. (1999). A note on superadditive probability judgment. *Psychological Review, 106*, 210–214. https://doi.org/10.1037/0033-295X.106.1.210

Macrae, C. N., Shepherd, J. W., & Milne, A. B. (1992). The effects of source credibility on the dilution of stereotype-based judgments. *Personality and Social Psychology Bulletin, 18,* 765–775. https://doi.org/10.1177/0146167292186013

McKenzie, C. R. M., Lee, S. M., & Chen, K. K. (2002). When negative evidence increases confidence: Change in belief after hearing two sides of a dispute. *Journal of Behavioral Decision Making*, *15*, 1–18. https://doi.org/10.1002/bdm.400

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*, 254–278. https://doi.org/10.1037/0033-295X.100.2.254

Meyvis, T., & Janiszewski, C. (2002). Consumers' beliefs about product benefits: The effect of obviously irrelevant product information. *Journal of Consumer Research, 28*, 618–635. http://dx.doi.org/10.1086/338205

Millroth, P., Guath, M., & Juslin, P. (2019). Memory and decision making: Effects of sequential presentation of probabilities and outcomes in risky prospects. *Journal of Experimental Psychology: General*, *148*(2), 304–324. https://doi.org/10.1037/xge0000438

Morey, R. D. & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes factors for common designs* (R package version 0.9.12-4.2) [Computer software]. https://CRAN.R-project.org/package=BayesFactor

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Nilsson, H., Olsson, H., & Juslin, P. (2005). The cognitive substrate of subjective probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 600–620. https://doi.org/10.1037/0278–7393.31.4.600

Nilsson, H., Rieskamp, J., & Jenny, M. A. (2013). Exploring the overestimation of conjunctive probabilities. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2013.00101

Nilsson, H., Winman, A., Juslin, P., & Hansson, G. (2009). Linda is not a bearded lady: Configural weighting and adding as the cause of extension errors. *Journal of Experimental Psychology: General*, *138*, 517–534. https://doi.org/10.1037/a0017351

Nisbett, R., Zukier, H., & Lemley, R. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology, 13,* 248–277. https://doi.org/10.1016/0010-0285(81)90010-4

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review, 7,* 375–402.

Peters, E., Dieckmann, N., Dixon, A., Hibbard, J. H., & Mertz, C. K. (2007). Less is more in presenting quality information to consumers. *Medical Care Research and Review, 64,* 169–190. https://doi.org/10.1177/1077558707064002030l

Peters, E., & Rothbart, M. (2000). Typicality can create, eliminate, and reverse the dilution effect. *Personality and Social Psychology Bulletin*, *26*, 177–187. https://doi.org/10.1177/0146167200264005

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117,* 864–901. https://doi.org/10.1037/a0019737

Plous, S. (1991). Biases in the assimilation of technological breakdowns: Do accidents make us safer? *Journal of Applied Social Psychology*, *21*(13), 1058–1082. https://doi.org/10.1111/j.1559-1816.1991.tb00459.x

Pothos, E. M., Barque-Duran, A., Yearsley, J. M., Trueblood, J. S., Busemeyer, J. R., & Hampton, J. A. (2015). Progress and current challenges with the quantum similarity model. *Frontiers in Psychology*, *6*, Article 205. https://doi.org/10.3389/fpsyg.2015.00205

Pothos, E. M., & Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behavioral and Brain Sciences*, *36*, 255–274. https://doi.org/10.1017/S0140525X12001525

Pothos, E. M., Busemeyer, J. R., & Trueblood, J. S. (2013). A quantum geometric model of similarity. *Psychological Review, 120*(3), 679–696. https://doi.org/10.1037/a0033142

Pothos, E. M., & Trueblood, J. S. (2015). Structured representations in a quantum probability model of similarity. *Journal of Mathematical Psychology*, 64, 35–43. https://doi.org/10.1016/j.jmp.2014.12.001

Read, D., & Grushka-Cockayne, Y. (2011). The similarity heuristic. *Journal of Behavioral Decision Making, 24*, 23–46. https://doi.org/10.1002/bdm.679

Roussel, J.-L., Fayol, M., & Barrouillet, P. (2002). Procedural vs. direct retrieval strategies in arithmetic: A comparison between additive and multiplicative problem solving. *European Journal of Cognitive Psychology*, *14*, 61–104. https://doi.org/10.1080/09541440042000115

Russo, J. E. (2015). The predecisional distortion of information. In E. A. Wilhelms & V. F. Reyna (Eds.), *Neuroeconomics, judgment, and decision making* (pp. 91–110). Psychology Press.

Russo, J. E., Meloy, M. G., & Medvec, V. H. (1998). Predecisional distortion of product information. *Journal of Marketing Research*, *35*, 438–452. http://dx.doi.org/10.2307/3152163

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893. https://doi.org/10.1016/j.tics.2016.10.003

Sanborn, A. N., Noguchi, T., Tripp, J., & Stewart, N. (2020). A dilution effect without dilution: When missing evidence, not non-diagnostic evidence, is judged inaccurately. *Cognition*, *196*, Article 104110. https://doi.org/10.1016/j.cognition.2019.104110

Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review*, *112*, 610–628. https://doi.org/10.1037/0033-295X.112.3.610

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464. https://doi.org/10.1214/aos/1176344136

Shanteau, J. C. (1970). An additive model for sequential decision making. *Journal of Experimental Psychology, 85*, 181–191. https://doi.org/10.1037/h0029552

Shanteau, J. C. (1972). Descriptive versus normative models of sequential inference judgment. *Journal of Experimental Psychology, 93,* 63–68. https://doi.org/10.1037/h0032509

Shanteau, J. C. (1975). Averaging versus multiplying combination rules of inference judgment. *Acta Psychologica, 39,* 83–89. https://doi.org/10.1016/0001-6918(75)90023-2

Shelton, S. (1999). The effect of experience on the use of irrelevant evidence in auditor judgment. *Accounting Review, 74*, 217–224. https://doi.org/10.2308/accr.1999.74.2.217

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323. https://doi.org/10.1126/science.3629243

Smith, H. D., Stasson, M. F., & Hawkes, W. G. (1999). Dilution in legal decision making: Effect of non-diagnostic information in relation to amount of diagnostic evidence. *Current Psychology, 17*, 333–345. https://doi.org/10.1007/s12144-998-1015-6

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning*. MIT Press.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318. https://doi.org/10.1016/j.tics.2006.05.009

Tentori, K., & Crupi, V. (2012). How the conjunction fallacy is tied to probabilistic confirmation: Some remarks on Schupbach (2012). *Synthese, 184*, 3–12. https://doi.org/10.1007/s11229-009-9701-y

Tentori, K., Crupi, V., & Russo, S. (2013). On the determinants of the conjunction fallacy: Probability versus inductive confirmation. *Journal of Experimental Psychology: General, 142,* 235–255. http://dx.doi.org/10.1037/a0028770

Tetlock, P. E., & Boettger, R. (1989). Accountability: A social magnifier of the dilution effect. *Journal of Personality and Social Psychology*, *57*, 388–389. https://doi.org/10.1037/0022-3514.57.3.388

Tobler, P. N., O'Doherty, J. P., Dolan, R. J., & Schulz, W. (2005). Human neural learning depends on reward prediction errors in the blocking paradigm. *Journal of Neurophysiology*, *95*, 301–310. https://doi.org/10.1152/jn.00762.2005

Troutman, C. M., & Shanteau, J. (1977). Inferences based on nondiagnostic information. *Organizational Behavior and Human Performance, 19,* 43–55. https://doi.org/10.1016/0010-0285(81)90010-4

Trueblood, J. S., & Busemeyer, J. (2010). A comparison of the belief-adjustment model and the quantum inference model as explanations of order effects in human inference. In In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (pp. 1166–1171). Cognitive Science Society.

Trueblood, J. S., & Busemeyer, J. R. (2011). A quantum probability account of order effects in inference. *Cognitive Science, 35,* 1518–1552. https://doi.org/10.1111/j.1551-6709.2011.01197.x

Trueblood, J. S., & Pothos, E. M. (2014). A quantum probability approach to human causal reasoning. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 1616–1621). Cognitive Science Society.

Trueblood, J. S., Pothos, E. M., & Busemeyer, J. R. (2014). Quantum probability theory as a common framework for reasoning and similarity. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.00322/full

Trueblood, J. S., Yearsley, J. M., & Pothos, E. M. (2017). A quantum probability framework for human probabilistic inference. *Journal of Experimental Psychology: General*, *146*(9), 1307–1341. https://doi.org/10.1037/xge0000326

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327–352. https://doi.org/10.1037/0033-295X.84.4.327

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. https://doi.org/10.1037/h0031322

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315. https://doi.org/10.1037/0033-295X.90.4.293

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*, 547–567. https://doi.org/10.1037/0033-295X.101.4.547

von Helversen, B., Herzog, S. M., & Rieskamp, J. (2014). Haunted by a doppelgänger: Irrelevant facial similarity affects rule-based judgments. *Experimental Psychology*, *61*(1), 12–22. https://doi.org/10.1027/1618-3169/a000221

von Helversen, B., & Rieskamp, J. (2009). Models of quantitative estimations: Rule-based and exemplar-based processes compared. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 867–889. https://doi.org/10.1037/a0015501

Walker, L., Thibaut, J., & Andreoli, V. (1972). Order of presentation at trial. *Yale Law Journal, 82,* 216–226.

Waller, W. S., & Zimbelman, M. F. (2003). A cognitive footprint in archival data: Generalizing the dilution effect from laboratory to field settings. *Organizational Behavior and Human Decision Processes, 91,* 254–268. https://doi.org/10.1016/S0749-5978(03)00024-4

Wansink, B., Kent, R. J., & Hoch, S. J. (1998). An anchoring and adjustment model of purchase quantity decisions. *Journal of Marketing Research, 35*(1), 71–81 https://doi.org/10.2307/3151931

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*(3), 273–281. https://doi.org/10.1080/14640746808400161

White, L., Pothos, E. M., & Busemeyer, J. R. (2013). A quantum probability perspective on the nature of psychological uncertainty. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 1599–1604). Cognitive Science Society.

Table 1

Median Parameter Values of All Models in All Studies

Model	Parameter	1	2	3	4
Similarity updating	c	.22	.54	.38	.45
	θ	14.01	9.95	15.64	10.80
	τ	.57	.64	.50	.54
	σ	.14	.14	.11	.12
Bayesian	σ	.27	.27	.38	.33
PT+N	d	.21	.22	.27	.21
	σ	.18	.21	.17	.22
Baseline	σ	46,910.56	2,357	.41	.61

Note. PT+N = probability-theory-plus-noise model.

Table 2

The Models' Bayesian Information Criteria (BICs) for the Probability Judgments in All Studies

Study	Result	Model			
		Similarity updating	Bayesian	PT+N	Baseline
1	Median BIC	-237.10	-82.6	-165.08	4.45
	n	24	0	1	0
2	Median BIC	-240.0	-84.28	-144.95	4.45
	n	25	0	1	0
3	Median BIC	-405.81	-46.55	-244.74	-28.95
	n	21	0	3	0
4	Median BIC	-262.38	-55.04	-121.39	-2.67
	n	23	1	1	0

Note. PT+N = Probability-theory-plus-noise model.

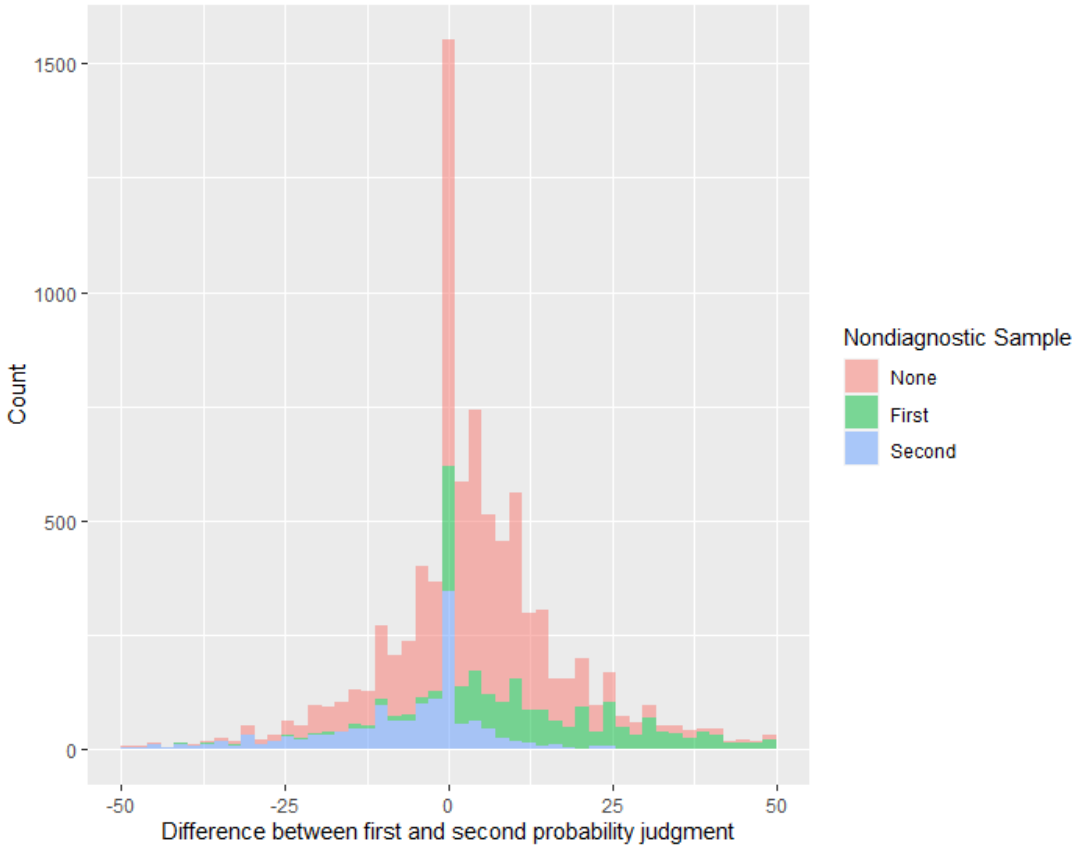


Figure 1. Histogram of differences between participants' first and second probability judgments by presentation of nondiagnostic sample in all trials in all four studies.

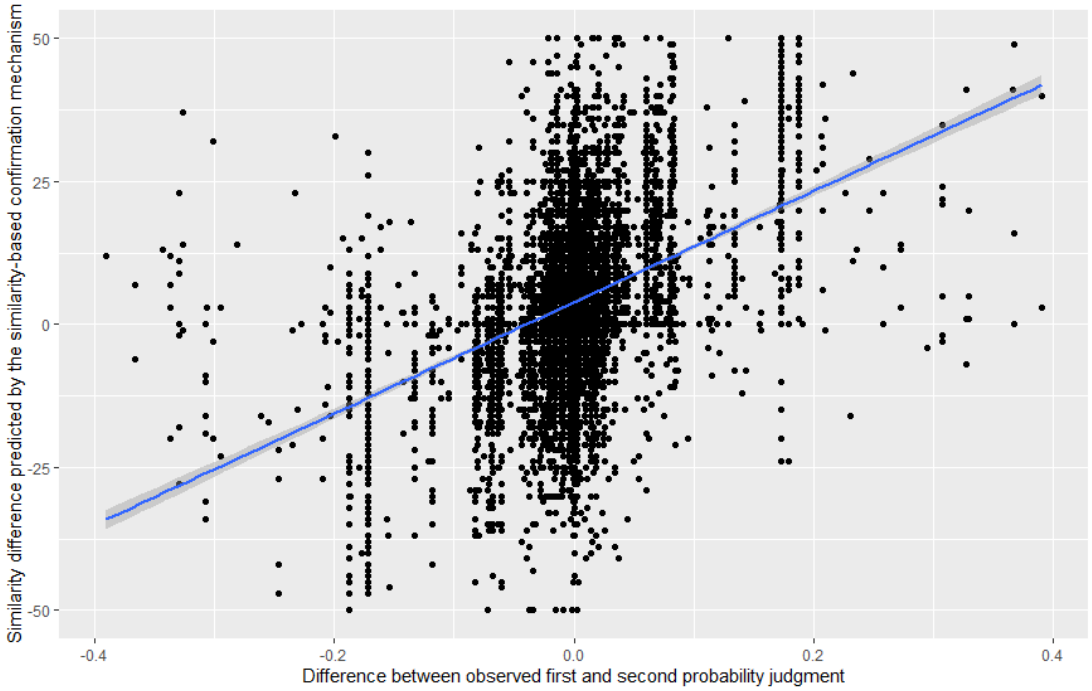


Figure 2. Correlation of the difference in participants' probability judgments in all four experiments between the first and the second judgment (combined probability estimate for Samples 1 and 2 minus the probability for Sample 1) and the similarity difference predicted by the similarity-updating model (Equation 1). Parameter c of the model used to produce the predictions is the mean value of the median parameter values over all experiments as reported in Table 2 ($c = .4$).

Appendix A

Detailed Implementation of the PT+N model

Our implementation of the PT+N model, in principle, follows the implementation of the Bayesian model introduced in Appendix B, with the exception that we assume that there is a chance that samples are perceived incorrectly, more specifically, a chance d that the color of a card is misperceived, leading to an incorrect count.

In our setting, the error chance d modulates how a sample of n cards is actually perceived. To model this, we first calculate the chance for every unordered sample of size N given the real sample E_1. For an arbitrary sample E_1' the probability of accidentally perceiving E_1 as E_1' given d is

$$P_{\text{perceive}}(E_1'|E_1, d) = \frac{N!}{dist(E_1,E_1')! \cdot \left(n-dist(E_1,E_1')\right)!} \cdot \left((1-d)^{N-dist(E_1,E_1')} \cdot d^{dist(E_1E_1')}\right) \quad \text{(A1)}$$

The distance between the two samples E_1 and $E_1{}'$is given by the number of different cards.

The likelihood that a sample E_1 stems from Deck A is calculated by multiplying the probability of perceiving every unordered sample, $P_{\text{perceive}}(E_1'|E_1, d)$, with the probability that this perceived sample stems from the deck, $P(E_1'|\text{A})$,

$$P(E_1|\text{A}) = \sum_{E_1'} P_{\text{perceive}}(E_1'|E_1, d) \cdot P(E_1'|\text{A}) \quad \text{(A2)}$$

The probability that the misperceived sample E_1' stems from Deck A is given by

$$p(E_1|\text{A}) = \frac{N!}{f_{1.1}! \times f_{1.2}! \times f_{1.3}!} \times p_{\text{A1}}^{f_{1.1}} \times p_{\text{A2}}^{f_{1.2}} \times p_{\text{A3}}^{f_{1.3}} \quad \text{(A3)}$$

The updating process is calculated analogous to the Bayesian cognitive model (described in Appendix B). Please note that the Bayesian cognitive model is nested in this implementation of the PT+N model ($d = 0$). The higher the error chance d, the higher are the deviations of the model's predictions from the predictions of the Bayesian cognitive model. However, the deviations are symmetrical and cannot in principle predict stimulus-dependant deviations such as the dilution effect.

Appendix B

Detailed Implementation of the Bayesian Model

The first step to calculate the probability of Hypothesis A given a specific piece of evidence is to calculate the likelihood of observing the sample under Hypothesis A. In our card game example, this is the likelihood of sampling the cards of Sample 1 out of Deck A:

$$p(E_1|\mathrm{A}) = p_{\mathrm{A1}}^{f_{1.1}} \times p_{\mathrm{A2}}^{f_{1.2}} \times p_{\mathrm{A3}}^{f_{1.3}}, \quad \text{(B1)}$$

where N is the sample size, p_{A1}, p_{A2}, and p_{A3} are the probabilities of the different colors in Deck A and $f_{1.1}$, $f_{1.2}$, and $f_{1.3}$, are the frequencies of the respective colors observed in Sample 1. The posterior probability of Sample 1 coming from Deck A is then computed by

$$p(\mathrm{A}|E_1) = \frac{p(E_1|\mathrm{A}) \times p(\mathrm{A})}{p(E_1|\mathrm{A}) \times p(\mathrm{A}) + p(E_1|\mathrm{B}) \times p(\mathrm{B})}, \quad \text{(B2)}$$

where $p(E_1|\mathrm{A})$ and $p(E_1|\mathrm{B})$ are the likelihoods of receiving Sample 1 out of Deck A and Deck B and $p(\mathrm{A})$ and $p(\mathrm{B})$ are the prior probabilities of Categories A and B, respectively. We implemented this model fixing the prior probabilities to .50, assuming that prior to having seen any data, participants would be indifferent about the categories. This posterior probability becomes a new prior probability in light of which additional information will be processed.

According to Bayesian theory, this new prior probability is updated in light of new evidence as follows by first computing the likelihood of observing Sample 2 out of Deck A:

$$p(E_2|\mathrm{A}) = p_{\mathrm{A1}}^{f_{2.1}} \times p_{\mathrm{A2}}^{f_{2.2}} \times p_{\mathrm{A3}}^{f_{2.3}}, \quad \text{(B3)}$$

where p_{A1}, p_{A2}, and p_{A3} are the probabilities of the different colors in Deck A and $f_{2.1}$, $f_{2.2}$, and $f_{2.3}$, are the frequencies of the respective colors observed in Sample 2. The posterior probability that both Sample 1 and Sample 2 come from Deck A is then computed by

$$p(\mathrm{A}|E_1, E_2) = \frac{p(\mathrm{A}|E_1)\times p(E_2|\mathrm{A})}{p(\mathrm{A}|E_1)\times p(E_2|\mathrm{A})+p(\mathrm{B}|E_1)\times p(E_2|\mathrm{B})}, \quad \text{(B4)}$$

where $p(\mathrm{A}|E_1)$ and $p(\mathrm{B}|E_1)$ are the posterior probabilities for Decks A and B given that Sample 1 was observed (and thus the new prior probabilities), and $p(E_2|\mathrm{A})$ and $p(E_2|\mathrm{B})$ are the likelihoods of receiving Sample 2 out of Deck A and B. Note that in this example, because the second sample is just as likely to come from Deck A as from Deck B, $p(\mathrm{A}|E_1,E_2) = p(\mathrm{A}|E_2)$.

Equation B4 can be rearranged to an odds format where

$$\frac{p(\mathrm{A}|E_1,E_2)}{p(\mathrm{B}|E_1,E_2)} = \frac{p(\mathrm{A})}{p(\mathrm{B})} \times \frac{p(E_1|\mathrm{A})}{p(E_1|\mathrm{B})} \times \frac{p(E_2|\mathrm{A})}{p(E_2|\mathrm{B})} = \frac{p(E_1|\mathrm{A})}{p(E_1|\mathrm{B})} \times \frac{p(E_2|\mathrm{A})}{p(E_2|\mathrm{B})}, \text{ if } p(\mathrm{A}) = p(\mathrm{B}) = .50. \quad \text{(B5)}$$

Appendix C

Behavior of Similarity-Based Confirmation

Figures C1 and C2 illustrate the anchoring-and-adjustment process.

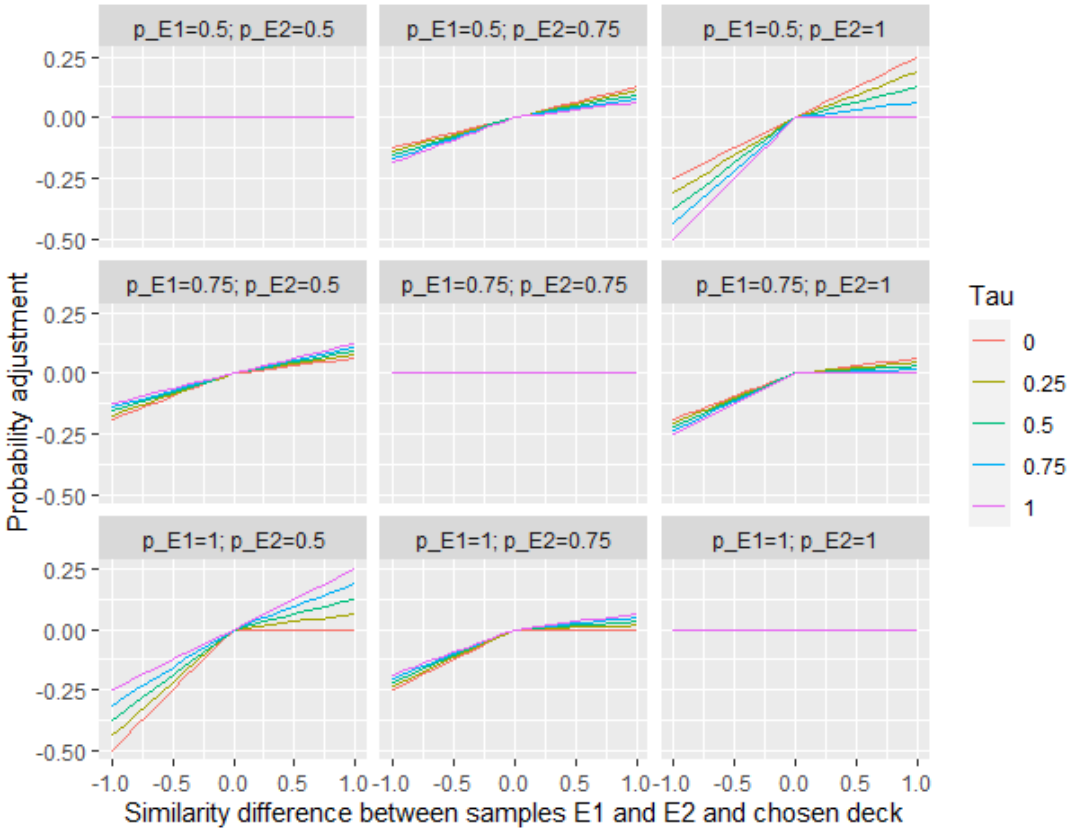


Figure C1. Adjustments predicted by the similarity-based confirmation mechanism for differences between the first samples E_1, E_2 and the chosen hypothesis, combinations of associated probabilities, and different values of τ.

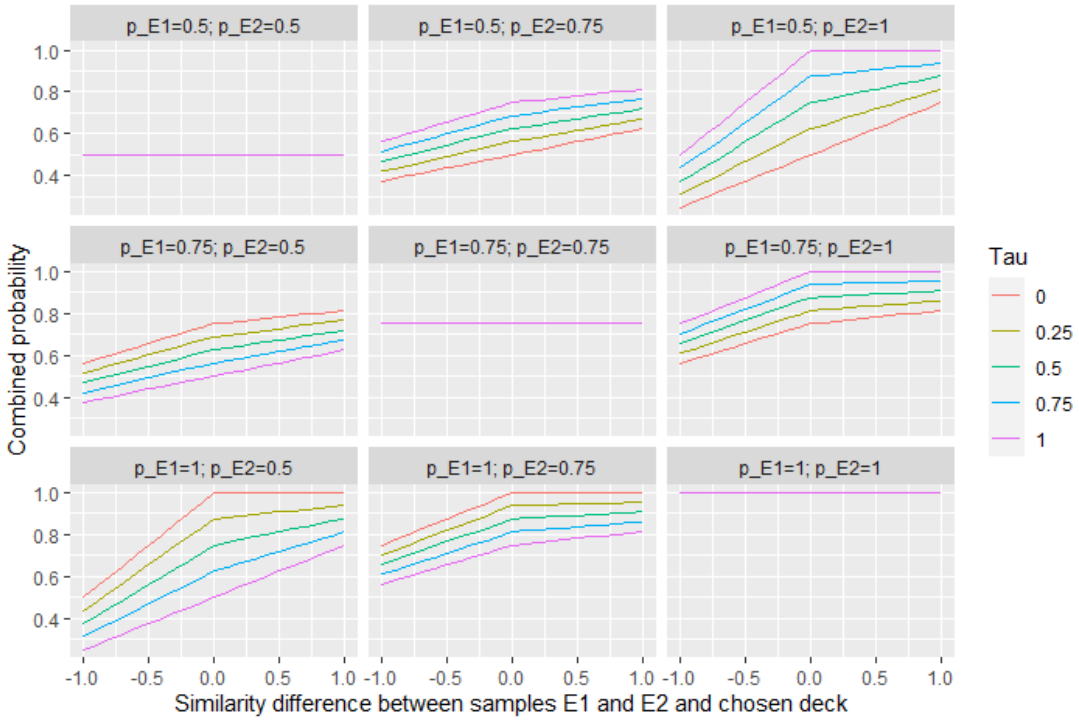


Figure C2. Combined probability predicted by the updating mechanism for differences between the first samples E_1, E_2 and the chosen hypothesis, combinations of associated probabilities, and different values of τ.

Appendix D

Comparison Between Likelihood and Similarity

Figure D1 shows the similarity (for different values of the sensitivity parameter *c*) and likelihood for a sample relative to the city-block distance between different samples and all tested decks. Each grid cell shows the results for another, representative sample. Grid cell 0-0-7, for example, shows the results (similarity/likelihood) for a sample with 0 blue, 0 green and 7 red cards relative to the distance between this sample and all possible tested decks. The presented sample types are representative for all samples, because city-block distance is symmetrical, meaning that the graphs for 0-0-7 and 0-7-0 are identical. The results show that both the likelihood and the similarity decrease with the distance between a sample and a deck. However, for small distances, the likelihood is much lower on average than similarity and consequently the decrease is flatter compared to similarity. Similarity can only approximate likelihood for some sample types and high values of the sensitivity parameter *c*.

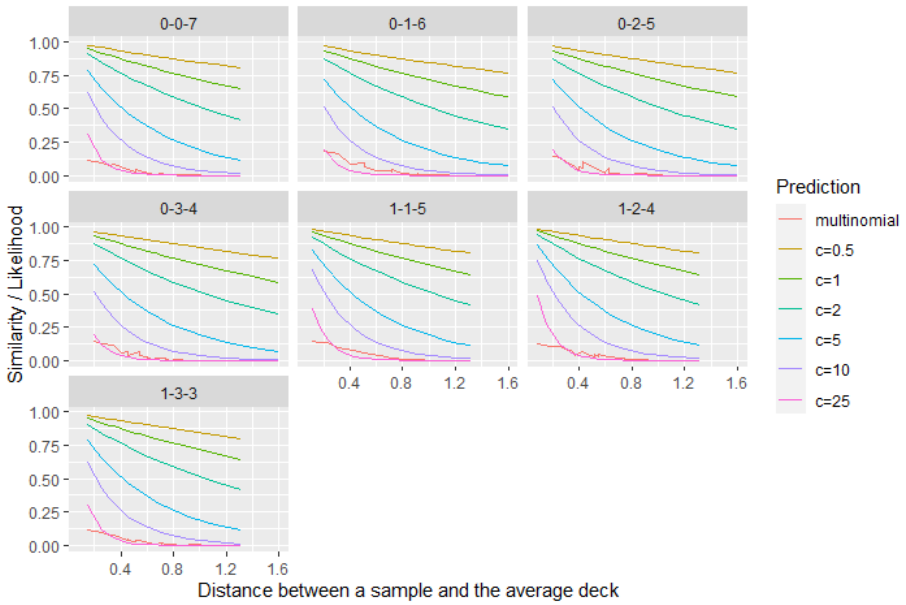


Figure D1. Comparison of the likelihood (calculated with the multinomial distribution) and similarities calculated with different values of parameter *c* relative to the distance between an average deck and different types of samples (grid).

Appendix E

Instructions of the Four Studies

In the following, you will be presented with two decks of cards. Both decks consist of 3 types of card: red, blue, and green cards. Each deck consists of 100 cards. The numbers above the decks indicate how many red, blue, and green cards the decks contain. Proceed by hitting any key.

One of the decks will be randomly picked and two [three] samples will be sequentially drawn (with replacement) from this deck. The cards that are drawn in each sample are replaced before drawing additional samples. Thus, within one game, the decks always consist of the number of cards indicated above the decks. The composition of the decks will vary between games. Proceed by hitting any key.

Your task will be to indicate which deck you think the two [three] samples were drawn from. Both samples were drawn from the same deck. Additionally, you will assess the probability that the samples stem from the deck you picked. **PLEASE NOTE THAT THE PROBABILITY JUDGMENTS ALWAYS RELATE TO THE DECK THAT YOU PICKED.** Proceed by hitting any key.

You will receive 15 CHF (or 2 course credits) for your participation. In the end, one of the games you played will be picked at random and played out. If you picked the right deck in this game, then you receive 2.50 CHF in addition to the participation fee as a bonus. Proceed by hitting any key.

Additionally, you can win a bonus, which is contingent on the accuracy of your probability judgment. The better your judgment, the higher your bonus will be (max. 5 CHF). Thus, you will receive a bonus for your choice AND your probability judgment. Proceed by hitting any key.

Please address the instructor now if anything is unclear. Note that you will always first pick a deck by hitting either key "1" or "2" and then provide your probability judgment. **PLEASE NOTE THAT WITHIN ONE GAME, BOTH [ALL THREE] SAMPLES WILL BE DRAWN FROM THE SAME DECK.**[E1] First, a couple of practice trials will follow, which will not count.[E2] Proceed by hitting any key.

***

Are you ready for the real experiment? Now every game counts.[E2] Please address the instructor if anything is unclear. If you are ready, then please hit any key.

Footnotes

[1] Note that this is true only if the pieces of evidence are independent.

[2] To test the ecological validity of the similarity-updating model we ran a simulation that parallels simulations by Juslin et al. (2009). This simulation was intended to test if the model leads to good judgments in our environment as compared to a normative solution as given by likelihood computation and Bayes's theorem. In sum, the similarity-updating model's binary predictions are well adapted to the environment and their accuracy increases with decreasing sampling error.

[3] Note that this definition of similarity assumes an equal number of features in i and j. Considering the task at hand, if we assume that sampling a certain feature (e.g., green cards) immediately brings one to reject a deck because it does not contain that feature, we need to (a) extend the missing features with an F (false) and (b) extend our notion of similarity:

$$s_{i,j}^{\text{ext}} = \begin{cases} s_{i,j} & \text{if } \forall \mathrm{A}_{i_k} \in i = (i_1, \dots, i_n), \forall j_k \in s = (j, \dots, j_n) \colon (i_k \wedge j_k) \\ 0 & \text{else} \end{cases}$$

The conjunction $(i \wedge j)$ becomes false if one conjunct is false. Intuitively, false in the present paradigm means that a sample s cannot have been drawn from a Deck A. In that case, the similarity is 0 and, thus, the probability for that deck is also 0.

[4] Note that Equation 4 can be reformulated in the tradition of reinforcement learning models $(p_t(\mathrm{A}) = p_{t-1}(\mathrm{A}) + \beta[p(\mathrm{A}) - p_{t-1}(\mathrm{A})])$ where β is the weighting parameter with values between 0 and 1, which indicates the weight on the difference between the two probability judgments—in other words, on the prediction error of the first judgment relative the second judgment, the latter being based on additional information (Sutton & Barto, 1998). Also note that if the weighting parameter takes the function of $\beta = 1/(n-1)$, Equation 4 produces the current or running mean over all previously encountered pieces of evidence.

[E1] This sentence was added to the instructions of Studies 2 and 3 to make sure participants thoroughly understood the task.

[E2] This sentence was used only in Studies 2 and 3.

Coordinating several mental strategies requires integration: Evidence from human judgment

Janina A. Hoffmann *[1,2], Rebecca Albrecht[3], and Bettina von Helversen[4,5]

[1]University of Bath
[2]University of Konstanz
[3]University of Basel
[4]University of Bremen
[5]University of Zürich

August 7, 2020†

Abstract

Individuals solve many real-world problems by applying distinct mental strategies, requiring them to find out, which strategy solves the current problem best. How people develop preferences for strategies and coordinate their use has, however, remained largely unresolved. Many theories of the mind postulate that individuals apply only one strategy at a given time, implying shifting between strategies depending upon domain-specific knowledge. On the flip side, integrating or blending knowledge from several strategies often proves beneficial. We argue here that a principled learning account provides valuable insights into how people solve the strategy selection problem. We present a generalized learning model of strategy selection and coordination, which reveals how assumptions about strategy coordination (i.e. switching between strategies or blending strategy knowledge) constrain strategy selection learning. We demonstrate that developing domain-specific strategies is only possible if people keep more than one strategy active at any given point in time, ruling out trial-by-trial strategy shifts. Two empirical experiments in which we vary the amount of context knowledge people can acquire support the conclusion that learning to rely upon several judgment strategies is better predicted by strategy-blending than strategy-shifting. Learning models may thus provide a suitable tool for understanding the basics of strategy coordination.

Keywords— Strategy selection, Reinforcement learning, Judgment

*[2]Janina A. Hoffmann, Department of Psychology, University of Bath, BA2 7AY, Bath, United Kingdom. E-mail: j.a.hoffmann@bath.ac.uk

†JAH, RA, BvH designed the research and wrote the manuscript. JAH conceptualized the cognitive model, conducted the simulations, and analyzed the data. RA implemented the experiments.

The authors declare no conflict of interest. This research has been supported by the German Research Foundation grant no. HO 5815/1-1 to the first author.

In many personal and professional decision situations, people have to adapt their decision behavior to changing environmental influences. Imagine, for instance, a trader at the rise of the New Economy. During this period, new companies such as Google or Yahoo entered the stock market, many of them violating previously successful business models. Potentially, the trader identifies these companies as deviating from previously traded enterprises and establishes a new policy to evaluate the companies' future success. Alternatively, the trader may slowly adjust her current trading policy by incorporating predictors of success for these new companies into her trading policy. Changing task demands, such as new competitors on the stock market, thus challenge current policies and raise the need to adjust these policies or select an alternative one.

The general question of how people adjust and shift between distinct mental strategies in response to task demands and personal preferences has been coined the *strategy selection problem* [6, 33, 42, 44, 45, 53, 54, 56]. The strategy selection problem arises from the often held assumption within psychological science that people solve a task by applying two, several, or a multitude of distinct strategies. For instance, dual-process theories of reasoning posit that people sometimes solve a reasoning problem by applying a fast and associative heuristic process (System 1) and sometimes by applying a slow and analytic process (System 2) [13, 63]. Theories in language have emphasized that people often follow grammatical rules, but store and retrieve exceptions to those rules from memory [22, 59].

Across these domains, the multiple strategy assumption poses two sub-problems: a) When do people apply which strategy? Under time pressure, for instance, the trader may promptly classify any new competitor as high-risk and refuse to invest, but under less pressing conditions the trader potentially considers additional factors that turn his decision around. Task and strategy demands like this, but also personal capabilities and strategy success shape which strategy a person follows [13, 27, 46, 53]. b) How do people coordinate different strategies? For one, the trader may shift between strategies and infer a new companies' future success sometimes by comparing it to similar ones and at other times by evaluating objective performance indicators. Alternatively, the trader may combine both strategies, that is, simultaneously compare the company to its competitors and evaluate it on objective criteria. *Strategy coordination* refers to the problem of how people combine and integrate several strategies [56].

Most psychological research has focused on the first problem to explain people's behavior, but neglected the question of how people coordinate different strategies. In this paper, we argue that to solve the strategy selection puzzle and to predict how people approach a given task, one needs to explicitly consider how people combine and coordinate these strategies. In particular, assumptions about strategy coordination constrain if people can —in principle —acquire general strategy preferences and can adapt strategies to domain-specific (or contextual) knowledge.

Here, we demonstrate analytically, in simulations, and empirically how strategy coordination impacts learning about the strategies and when to use them. Analytical results rule out a strict interpretation of strategy shifting, but still allow for blending strategy knowledge or selectively activating strategies depending on the domain-specific information. Using judgment research as an example, we explore in simulations how less extreme preferences for shifting and blending affect strategy choice. We then identify empirical conditions to contrast different types of strategy coordination and demonstrate empirically that strategy-blending best predicts human judgments.

1 Strategy coordination: Blending vs shifting

A range of prominent theories in psychological science, such as the adaptive toolbox approach, have postulated a *strategy shifting account* [13, 18, 19, 31, 63]: People select one strategy out of a set of competing strategies to solve the problem at hand and follow this single strategy, ignoring alternative solutions another strategy might yield. Yet, more recent evidence points towards the possibility that even in cases in

which one strategy clearly dominates, people still activate the alternative strategy and integrate or *blend* the responses from several strategies [7, 9, 24, 60, 67]. For instance, people retrieve previous knowledge from memory, even if they consistently apply one decision rule [21, 66]. Although strategy coordination is a recurring topic across domains in cognitive science from reasoning [13, 63] to language learning [22, 59], the question of whether individuals shift among or blend different strategies remains unresolved.

To illustrate strategy coordination, imagine each strategy as one out of several experts shouting out their advice. You can select the expert depending on their domain of knowledge and follow their solution, a shifting between experts, or integrate their solutions independent of their domain-specific expertise, a blending account. In the extremes, one expert provides the solution to one problem in only one domain in strategy-shifting or every expert provides a solution to all problems in all domains in strategy-blending. Thus, strategy-shifting implies that people heavily consider and develop contextual knowledge to select a strategy, whereas strategy-blending assume that domain-specific expertise is ignored and only general strategy preferences are identified.

Previous work oftentimes demonstrated a performance benefit of strategy-blending. Combining forecasts from different methods (or individuals) regularly outperforms the single best forecasting method in prediction tournaments and forecasting research [3, 4, 16, 40, 47, 48, 64], in particular if those methods make different systematic errors [39]. In the same way, blending the output from several mental strategies may reduce the impact associated with errors in strategy choice [23, 24, 67]. Despite its statistical advantage, strategy-blending has been refuted as a plausible mechanism within one mind [19, 34, 61]. In addition, evidence is rare that individuals deliberately adopt strategy-blending [24, 39, 47]. For instance, collaborative teams often seek to identify the more appropriate estimate instead of averaging their estimates [15, 41, 49, 50]. Such strategy shifting can be advantageous, too, namely, if strategies strongly differ in their accuracy [10, 38] or if averaging all opinions neglects novel or specialized knowledge from a minority [8, 55]. The latter argument hints at the possibility that a compromise between pure shifting and blending solves the strategy coordination problem best. Individuals may preferentially activate and integrate all strategies, but consider domain or context specific information for some problems to determine the most appropriate strategy for this subset (i.e. selective activation).

Taken together, although strategy-blending is statistically advantageous in most setups, this perspective has found little resonance in theories of the mind postulating strategy-shifting. In turn, how individuals prefer to coordinate strategy use does not necessarily coincide with the statistically optimal approach. One reason why conclusions may differ is that statistical aggregation approaches often limit their investigation to the question of how optimal policies are coordinated. Individuals, however, are usually confronted with ill-defined problems for which they have to *learn* appropriate strategies. Taking a learning perspective may resolve this divide between optimal strategy coordination and observed human preferences by uncovering the psychological limits of strategy coordination approaches.

2 The need for learning models to solve the strategy selection puzzle

The idea that strategies can be conceived as experts out of which the individual selects the best one for the current problem has dominated research on strategy selection in decision making [34, 56], memory [65], and reasoning [13, 17, 57] for a long time. Consequently, strategy-blending has been mostly contrasted with strategy-shifting under the assumption that strategies have already been adapted to the problem at hand [7, 23, 44]. Yet, this analogy falls short of a sufficient explanation for individual problem solving because during the learning process individuals need to simultaneously infer how to successfully execute each strategy *and* which one they should predominantly execute.[1]

[1]Some decision making models implement both learning processes, but do not explicitly expand on the interplay be-

Importantly, the degree to which people focus only on the most suitable strategy in each trial or engage in several strategies simultaneously constrains which knowledge individuals gain about each strategy and how likely one finds the best strategy over time. To illustrate this idea, reconsider the problem of advice taking. When following the advice of a single expert, this expert will be informed about the accuracy of her recommendation and can adjust her policy accordingly. The initially ignored experts, however, will neither know whether they would have made a correct prediction, nor will the decision maker be able to infer if the chosen expert provided the best solution. When integrating the advice from multiple experts, however, each expert will be updated about the correctness of the overall advice and adjust her recommendation strategy depending upon her initial contribution. Likewise, the decision maker can determine which expert overall provided a better solution, but is unable to tie it to domain-specific expertise.

Over a longer learning period, this interplay between strategy-learning and learning to coordinate strategy use predicts patterns of strategic preferences that are hard to anticipate without learning. For instance, multi-modal response distributions, that is the same person answering a question sometimes in one way and sometimes in a different way, have often been taken as evidence for strategy shifting [24, 34, 67]. Yet, strategy-blending also allows for multi-modal patterns to emerge if the integrated prediction from all strategies is not fused into a single response, but activates multiple response options out of which one is selected. Even more surprising, multi-modal response patterns should be rarely observed in strategy-shifting because only the currently activated strategy is adjusted and one does not learn to perceive multiple strategies as equally appropriate[2]. Thus, strategy predictions can be easily confounded with response selection processes if learning processes are ignored. This highlights that without a principled learning model it is difficult to distinguish between strategy-shifting and strategy-blending.

tween learning each strategy and learning to coordinate their use [34, 37]

[2]Indeed, the chosen strategy has to predict multi-modal response patterns to be able to observe those pattern in strategy-shifting.

3 Blending and shifting through the lens of a mixture-of-experts learning model

The idea that experts provide different solutions and one only needs to choose between or integrate these solutions is the key concept underlying mixture-of-experts models (Figure 1, see Appendix A). We delineate here a mixture-of-experts model for strategy selection, BASICS, that describes **B**lending **A**nd **S**hifting **I**n **C**oordinating **S**trategies. Generally speaking, mixture-of-experts models consist of two mechanisms: the experts, here the strategies $(s_1, ..., s_m)$, and the coordination mechanism, here the strategy selection mechanism. In the tradition of mixture-of expert models, BASICS efficiently learns which strategy solves a subset of problems best [29, 30]. Each strategy proposes for each possible solution $(c_1, ..., c_n)$ how likely it solves a given problem with dimensions $(x_1, ..., x_k)$, denoted by the choice probability $P_{c,s}$. The strategies can encompass simple solutions, such as heuristics based on a single cue or dimension [29], more intuitive approaches, such as a similarity-based comparisons to previous experiences [37], or sophisticated problem-solving strategies, such as a multidimensional non-linear integration of several dimensions [34], with each strategy defined as a function over the problem dimensions.

The strategy selection mechanism explicitly describes how several strategies are coordinated. In this mechanism, the probability of choosing solution c, the choice probability P_c, aggregates the probabilities of choosing solution c under each strategy s by weighing the choice probability for each strategy $P_{c,s}$ with the strategy activation a_s: $P_c = \sum_s a_s \cdot P_{c,s}$. This strategy activation, varying from 0 to 1, expresses strategy-shifting if a_s takes only the values 0 or 1 and thus only a single strategy is activated for solving the current problem. The strategy activation expresses strategy-blending if a_s takes any value between 0 or

1 and thus several strategies are simultaneously activated. How strongly each strategy is activated depends on two factors: global strategy preferences β_s and learned contextual knowledge about when to apply each strategy, $\sum_h a_h \cdot w_{s,h}$.

$$a_s = \frac{e^{\gamma * \sum_h a_h * w_{s,h} + \beta_s}}{\sum_s e^{\gamma * \sum_h a_h * w_{s,h} + \beta_s}} \quad (1)$$

The strategy selectivity γ modulates the degree to which contextual knowledge activates one strategy and ignores global strategy preferences causing strategy-shifting, or —on the contrary —contextual knowledge is neglected and global preferences dominate strategy activation, eventually causing strategy-blending (Figure 1, Appendix A.2). Under an indefinitely high strategy selectivity (dashed black line, $\gamma = 10000$), only a single strategy is activated to solve the current problem, namely the one preferred by contextual knowledge.[3] The steep transition from 0 to 1 in strategy activation clearly marks the all-or-none strategy choice in strategy-shifting. Under a very low strategy selectivity (dashed white line, $\gamma = .00001$), all strategies are activated depending upon the general preference for strategy s and the number the competing strategies, but any contextual knowledge is disregarded. The flat line marks this context-independent strategy-blending. At intermediate levels of strategy selectivity (e.g. black line, $\gamma = 1$), each strategy is activated depending upon global preferences and contextual knowledge, allowing for *selective activation* of one strategy, that is to preferably rely on one strategy, but to pick an alternative strategy to solve specific problems. Varying strategy selectivity thus allows to compare different accounts of strategy coordination during learning.

[3]The datasets generated and/or analysed during the current study are available in the Open Science Framework along with the code and experimental material, `https://osf.io/3hg27/?view_only=9f653f00832c47a0b65fcb06938395eb`

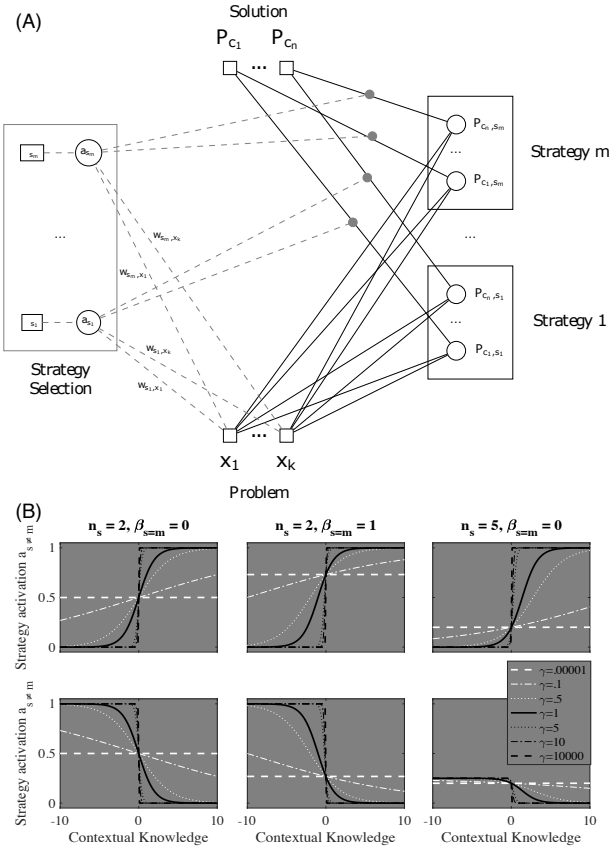


Figure 1: (A) Architecture of BASICS. Each strategy $s_1, .., s_m$ chooses each solution $c_1, ..., c_n$ to problem x with dimensions $x_1, .., x_k$ with choice probabilities $P_{c_1,s_1}, .., P_{c_n,s_m}$. These choice probabilities are integrated into overall choice probabilities for each solution $P_{c_1}, .., P_{c_k}$ depending upon the strategy activation of each strategy $a_{s_1}, ..., a_{s_m}$. (B) Strategy activation for a strategy $s = s_m$ (upper row) and the competing strategies (lower row) as a function of contextual knowledge about strategy s, strategy selectivity γ (separate lines), general strategy preference β (first two columns), and number of strategies (right column). Strategy activation for strategy s_m increases as a function of associated contextual knowledge and the general strategy preference, but declines with the number of available strategies. The strategy selectivity determines how sharply people transition from using one strategy to another, modulating the degree of strategy-blending to strategy-shifting.

During learning, BASICS learns simultaneously how and when to apply each strategy via trial-by-trial feedback. Strategy coordination, blending or shifting, directly impacts both aspects: learning of each strategy and learning of contextual knowledge. First, strategies are adjusted proportionally to their contribution. In blending, the strategies contribute relative to their strategy activation and are adjusted accordingly. In pure strategy-shifting, however, the activation of all but one strategy are zero and thus only one strategy can learn anything —the currently used one. Second, general strategy preferences β_s and context weights $w_{s,h}$ are adjusted as a function of strategy activation a_s, the context activation a_h, the strategy selectivity γ, the error attributed to each strategy E_s, and the combined error over all strategies E, $\Delta\beta_s = \lambda_s a_s(E - E_s)$ and $\Delta w_{s,h} = \lambda_s \gamma a_h a_s(E - E_s)$. Extreme values of strategy selectivity γ hinder the learning of contextual knowledge modelled by the context weights. In pure strategy-blending the low selectivity γ leads to a $\Delta w_{s,h}$ of near zero, implying that in strategy-blending people do not update any context dependent knowledge, but only a general strategy preference. In pure strategy-shifting the mixed error E reduces to the error elicited by the selected strategy E_s, i.e. $E = E_s$, implying that neither context-dependent knowledge, $\Delta w_{s,h} = 0$, nor general strategy preferences are adjusted. As a result, shifting between strategies in each single trial only enforces the strategy that appears most successful so far and does not help, in its extremes, to acquire knowledge about when to apply an alternative solution.

4 Strategy coordination in human judgment

Forecasting the nature of unknown future events is a prime example of strategy coordination in everyday life. Politicians, for instance, attempt to predict the public's opinion towards a policy by recalling their opinion towards similar historical events or by explicitly balancing the needs of different groups of voters. These forecasting, or judgment, tasks require the individual to combine multiple sources of information, such as the needs of different voters, into a quantitative prediction, such as the success of a policy, by following a suitable forecasting strategy. Rule-based strategies explicitly weigh different pieces of information by their importance to form a global judgment, whereas memory-based strategies compare the current information to similar past experiences. Individuals adapt these strategies to the problem at hand, their skills, and situational demands [27, 28, 32, 52], but how they coordinate these strategies remains an unresolved question [1, 7]. Further, such numeric prediction tasks facilitate to distinguish between strategy-blending and shifting because of the finer resolution of the response scale. Therefore, the domain of human judgment provides a well-suited test bed to investigate strategy coordination.

To understand strategy coordination within the domain of judgment, BASICS assumes that people base their judgment at any point in time on two strategies (Appendix B): a rule-based and a memory-based strategy. The rule-based strategy learns to abstract a linear, additive rule, whereas the memory-based strategy associates past instances, exemplars, with previous judgments and retrieves those exemplars based upon their similarity [31]. Over trials, BASICS develops a preference for a memory-based over a rule-based strategy, thereby distinguishing between global and contextual strategy preferences. This combination of global and contextual strategy preferences allows BASICS to store for which problems to retrieve past exemplars as an exception to the rule-based strategy, or, alternatively, to apply a rule as an exception to the memory-based strategy [12, 58]

4.1 Model Validation

We validated the psychological plausibility of BASICS by replicating important findings in the judgment literature (see Supplemental Information for model training): a) the finding that nonlinear functions are learned more slowly than linear ones [35] and b) more likely picked up by a memory-based strategy than by a rule-based strategy [27, 28, 31, 36]. To replicate these findings, BASICS predicted how fast people learn to solve four judgment tasks, two

linear and two non-linear ones [28, Figure 2]. Corresponding to empirical findings, judgment error decreased in all tasks with training and decreased faster for linear functions compared to non-linear ones. BASICS developed a clear preference for the rule-based strategy in linear tasks, but preferred the memory-based strategy in nonlinear tasks. Additional simulations suggested that BASICS was able to predict judgments for new objects outside the training range in linear tasks, but this extrapolation posed a problem in nonlinear tasks, replicating a frequently found pattern [31].

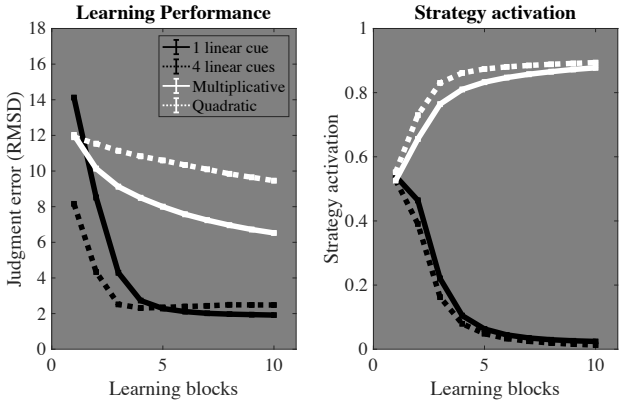


Figure 2: Predictive judgment error (left panel) and strategy activation of the memory-based strategy in four standard judgment tasks: a linear judgment task with one predictive cue, a linear task with four predictive cues with different cue weights, a multiplicative task including the interaction between several cues, and a quadratic task (separate lines). Judgment error, measured in RMSD between model-predicted judgments and the criterion, declines with the number of learning blocks. Nonlinear tasks preferably activate the memory-based strategy, whereas linear tasks do not.

Yet, the distinction between linear and non-linear tasks does not help to disentangle varying degrees of strategy-blending and strategy-shifting. Independent of the strategy selectivity γ, BASICS preferably picks up nonlinear tasks by adopting a memory-based strategy in our simulations, whereas it adopts a rule-based strategy in linear tasks. This suggests that previous work was not well-suited to answer the question of whether people blend different strategies or shift between them in single trials. To effectively investigate strategy coordination, it is necessary to study judgment tasks in which people have to apply several strategies (Supplemental Information). Rule-plus-exception tasks provide such a classical example [34, 51, 59].

4.2 Rule-plus-exception judgments

Accurately solving rule-plus-exception tasks requires participants to apply several judgment policies depending upon the item to judge. In our rule-plus-exception task, most items adhered to a linear additive rule, but some items represented exceptions contradicting the rule. Identifying and judging these exceptions should thus demand retrieving the exceptions from memory, whereas items following the linear rule should be judged according to a rule-based strategy.

In two experiments, we trained BASICS and human participants on the same rule-plus exception task varying between experiments how easily the exceptions can be identified (the degree of contextual knowledge) and within each experiment the frequency with which the exception items are repeated during training (10 or 50 exceptions, see Appendix C for details). After this initial learning phase, strategy-blending, selective activation, and strategy-shifting make distinct predictions about which items participants should judge accurately in a later test (judgment error) and which items they perceive as more familiar (familiarity-based choice, Figure 3 for a priori quantitative predictions and participant's judgments). In strategy-blending, how strongly individuals activate the rule-based and the memory-based strategy is independent of the item to judge, causing a high error on exceptions because the rule-based strategy is more strongly activated than the memory-based strategy. In strategy-shifting, in comparison, individuals should activate strategies in response to (item specific) contextual knowledge, advantaging the activation of a memory-based strategy for both exception items and all rule-following items.

If exceptions and rule-following items are repeated with a high frequency during training, individuals should judge exceptions as accurately as the repeated rule-following items. However, if strategies are selectively activated, developing contextual knowledge is only required to remember the exceptions, but not to judge any rule-following item. Thus, despite a similar accuracy for exceptions and rules, only exceptions will appear familiar.

We tested these predictions in two experiments varying the degree to which participants could gather contextual knowledge regarding specific items. In both experiments, individuals learned to judge random rule-following items, exceptions, and repeated rule-following items. In Experiment 1, participants learned to judge the same two exception items repeatedly (and two repeated rule-following items) to examine under conditions of high contextual knowledge whether individuals use this knowledge to switch between strategies. Experiment 2 generalizes the findings to a situation in which individuals only gather a gist of contextual knowledge by training participants on exception items (and rule-following items) that closely resemble each other but were never repeated during training.

Experiment 1: High contextual knowledge Overall, participants learned to make consistent judgments in the training phase (low frequency condition: $r = 0.69$, $SD = 0.19$; high frequency condition $r = 0.52$, $SD = 0.27$). The results during test provide a strong argument for strategy-blending compared to strategy-shifting or selective activation (see Figure 3 and Table 1). First, participants judged the exceptions less accurately than rule-following items that were presented as often as the exceptions during training. Second, how closely the test item resembled the trained item only mattered for exceptions presented with a high frequency and not for exceptions presented at a low frequency or rule-following items. This pattern of results matched the qualitative predictions of strategy-blending far better than the predictions of strategy-shifting. Familiarity-based choices indicated that participants built up some memory for exception and rule-following items during training in the high frequency condition and considered these items as more familiar than new items, in line with strategy-blending and strategy-shifting. Yet, participants did not judge exceptions as more familiar than rule-following items —a result clearly arguing against a selective activation of strategies.

To quantify the predictive accuracy of the models (see Table 1), we calculated the root mean square deviation (RMSD) between each model's predictions and each participant's judgments. On average, all learning models outperformed the baseline model, a model guessing the training mean, two-sided paired t-tests, all $t(102) < -10$, Cohen's $d < -1.4$, $p < .001$. Again, we found an advantage of blending over shifting for judgment. The blending model predicted judgments better than the shifting model, two-sided paired t-test, $t(102) = -4.6$, Cohen's d = -0.45, $p < .001$, as did selective activation, two-sided paired t-test, $t(102) = -5.1$, Cohen's d = -0.50, $p < .001$. As expected, blending and selective activation could not be discriminated based on the participants' judgments, two-sided paired t-test, $t(102) = -0.1$, Cohen's d = -0.02, $p < .849$. [4] For familiarity choices however, all models had difficulties to predict participants' choices because participants and the models were mostly guessing. A guessing model, predicting 50% chance of guessing old, fared better than the shifting model, two-sided paired t-test, $t(102) = -2.0$, Cohen's d = -0.20, $p = .043$, or the selective activation model, two-sided paired t-test, $t(102) = -3.5$, Cohen's d = -0.34, $p < .001$. The blending model, however, slightly predicted familiarity-based choices better than the guessing model, two-sided paired t-test, $t(102) = -2.9$, Cohen's d = -0.29, $p = .004$, showing a slight advantage for blending. Taken together, those results indicate that in situations in which people can easily build up contextual knowledge and select between strategies using this knowledge, they appear to preferably integrate or blend several strategies independent of the knowledge gathered.

[4] We also considered the correlation and the standardized mean absolute error. Both indicators reach the same conclusions.

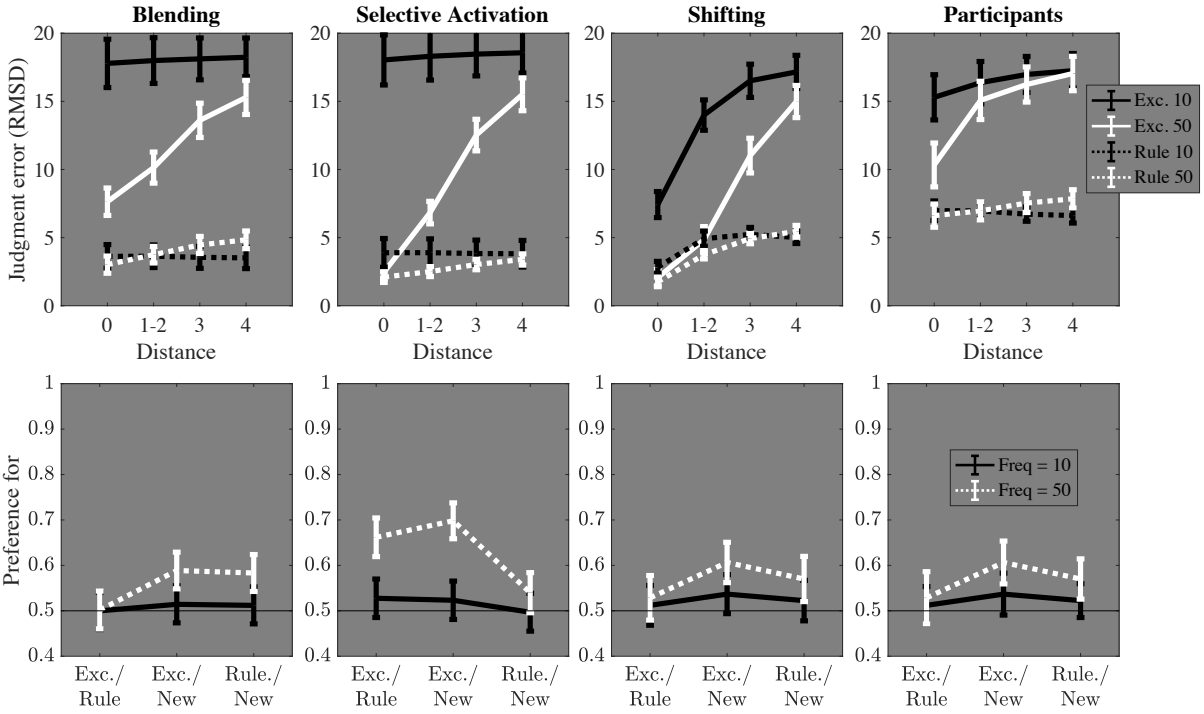


Figure 3: BASICS predictions for judgment error (upper panels) and familiarity-based choices (lower panels) as well as participant's judgment error (right upper panel) and participants' familiarity-based choices (right lower panel) for exceptions (Exc.), rule-following items (Rule), and new, unseen items (New). Participants (and the model) encountered the same two exception items (and rule-following items) 5 (black lines) or 25 times (white lines) during training. Model predictions are shown for strategy-blending (left panel, $\gamma = 0.10$), selective activation ($\gamma = 1$), and strategy-shifting ($\gamma = 10$). In test, participants judged the same items encountered in training (Distance = 0), but also items closely resembling the training items (Distance = 1-4, distance 1 and 2 were collapsed to one data point). Judgment error was measured as the RMSD between the correct criterion in training and predicted judgments in test. Familiarity-based choices depict how often participants chose one item type (e.g. the exception item) as more familiar than another item type (e.g. rule following item. Error bars indicate 95 % confidence intervals.

Experiment 2: Low contextual knowledge Experiment 2 aimed to generalize the findings to a more naturalistic situation in which individuals do not encounter exactly the same problem repeatedly, but have to solve highly similar problems, thus limiting the amount of contextual knowledge people can rely on. We implemented this idea in our experiments by asking participants to judge not the same exceptions in training, but exception items highly similar to each other. Overall, this study replicated the results from experiment 1. Participants made more errors on the exceptions than on the repeated rule items, independent of the frequency of repetitions during training, and did not judge the exceptions as more familiar than the repeated rule items. Predictive model accuracy likewise suggested that strategy-blending better explained participants' judgments and familiarity-based choices than strategy-shifting.

Table 1: Judgment error (in RMSD) for different types of items and predictive accuracy (in RMSD) for different models in Experiment 1 and 2.

	Judgment Error			
	Experiment 1		Experiment 2	
	Low Freq. (N = 51)	High Freq. (N = 52)	Low Freq. (N = 50)	High Freq. (N = 52)
Rule	6.9 (1.7)	7.4 (2.2)	6.1 (2.1)	7.5 (2.5)
Exception	16.7 (4.8)	15.4 (4.0)	17.4 (7.4)	13.4 (5.8)
Random	6.5 (1.5)	8.1 (2.0)	6.4 (1.8)	8.3 (2.7)
New	6.8 (1.6)	8.3 (2.0)	6.5 (2.3)	8.1 (3.8)
	Predictive accuracy			
	Judgment	Familiarity	Judgment	Familiarity
Baseline	10.3 (1.4)	.31 (.03)	9.4 (1.6)	0.17 (.07)
Blending	7.6 (2.0)	.30 (.03)	7.4 (2.4)	.16 (.07)
Shifting	8.3 (1.7)	.31 (.04)	7.8 (2.1)	.16 (.07)
Selective Activation	7.6 (1.8)	.32 (.04)	7.6 (2.3)	.19 (.07)

Note. RMSD = root mean square deviation between correct and participants' judgments (judgment error) or model-predicted and participants' judgments (predictive accuracy). Standard deviations in brackets.

5 General Discussion

The ability to adapt ones' strategies to current task demands is crucial in an uncertain and changing world. In this paper, we investigated the question of how people coordinate the use of several strategies applying a novel learning focus. Drawing inferences from a generalized mixture-of-experts model for strategy selection, we first demonstrated how strategy coordination in a single trial impacts two facets of strategic knowledge: how each strategy is adjusted and when it should be applied. This analysis scrutinized the often-held assumption that people select one strategy based upon domain-specific knowledge in a single trial, disregarding alternative solutions. Instantiating BASICS in the domain of human judgment next highlighted that the model replicates basic findings from the judgment literature and, even more importantly, allows to forecast decision making in new experiments. We used its predictive ability to tease apart three forms of strategy-coordination: strategy-shifting, strategy-blending, and selective activation. Overall, these experiments support the notion that individuals blend several strategies in a single trial [7, 9, 24, 60, 67], instead of choosing a single strategy [18, 19, 31, 63] or selectively activating strategies depending upon contextual knowledge [31, 34, 57].

Strategy selection research has regularly emphasized the notion that humans adapt strategies to the environment [5, 11, 28, 31, 53], mental resources [27, 43, 46, 57], or external demands [2, 26]. We found little evidence that individuals switched between strategies on a trial-by-trial level, as prominently proposed by influential theories of the mind [13, 19]. Importantly, our analysis ruled out strategy-switching as a psychologically well-founded explanation for strategy learning. Our experiments neither suggested that individuals selectively recruited alternative strategies to the predominant one depending upon domain-specific knowledge, as proposed by selective activation. Instead, we identified a surprising advantage for strategy-blending. One likely reason is that individuals had to simultaneously gather knowledge about each strategy and find out how to coordinate their use —a situation in which blending several strategies proves statistically beneficial [24]. Understanding from a trial-by-trial learning perspective how strategy coordination interacts with strategic knowledge thus helps to fine-tune and challenge often-held as-

sumptions in psychological theories. This focus may prove informative for areas beyond judgment and decision making, such as reasoning [63], memory [65], social cognition [20], or skill acquisition [62].

On first glance, the mixture-of-experts approach to strategy coordination may seem overwhelmingly resource-demanding as individuals need to execute all strategies at the same time. The strategies considered here, however, draw upon very different cognitive resources and thus a concurrent application may come at little cognitive costs [14, 28]. Alternatively, one can integrate cost-benefit analyses into the strategy coordination learning [42]. Future work may thereby explore how different forms of contextual knowledge, or domain-specific expertise, help to quickly adapt response strategies to real-world demands.

References

1. Albrecht, R., Hoffmann, J. A., Pleskac, T. J., Rieskamp, J. & von Helversen, B. Competitive retrieval strategy causes multimodal response distributions in multiple-cue judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition* (2019).
2. Allred, S., Duffy, S. & Smith, J. Cognitive load and strategic sophistication. *Journal of Economic Behavior & Organization* **125,** 162–178 (2016).
3. Armstrong, J. S. in *Principles of Forecasting: A Handbook for Researchers and Practitioners* (ed Armstrong, J. S.) (Kluwer Academic Publishers, Norwell, MA, 2001).
4. Atanasov, P. *et al.* Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management science* **63,** 691–706 (2017).
5. Beach, L. R. & Mitchell, T. R. A contingency model for the selection of decision strategies. *Academy of management review* **3,** 439–449 (1978).
6. Boureau, Y.-L., Sokol-Hessner, P. & Daw, N. D. Deciding how to decide: Self-control and meta-decision making. *Trends in cognitive sciences* **19,** 700–710 (2015).
7. Bröder, A., Gräf, M. & Kieslich, P. J. Measuring the relative contributions of rule-based and exemplar-based processes in judgment: Validation of a simple model. *Judgment and Decision Making* **12,** 491 (2017).
8. Budescu, D. V. & Chen, E. Identifying Expertise to Extract the Wisdom of Crowds. *Management Science* **61,** 267–280. ISSN: 0025-1909 (2015).
9. Chapman, G. B. & Johnson, E. J. Incorporating the irrelevant: Anchors in judgments of belief and value. *Heuristics and biases: The psychology of intuitive judgment,* 120–138 (2002).
10. Davis-Stober, C. P., Budescu, D. V., Dana, J. & Broomell, S. B. When is a crowd wise? *Decision* **1,** 79–101. ISSN: 23259973 (2014).
11. Dieussaert, K., Schaeken, W., Schroyens, W. & d'Ydewalle, G. Strategies during complex conditional inferences. *Thinking & reasoning* **6,** 125–160 (2000).
12. Erickson, M. A. & Kruschke, J. K. Rules and exemplars in category learning. *Journal of Experimental Psychology: General* **127,** 107–140. ISSN: 1939-2222 (June 1998).
13. Evans, J. S. B. *Hypothetical thinking: Dual processes in reasoning and judgement* (Psychology Press, 2019).
14. Fechner, H. B., Schooler, L. J. & Pachur, T. Cognitive costs of decision-making strategies: A resource demand decomposition analysis with a cognitive architecture. *Cognition* **170,** 102–122 (2018).
15. Floyd, R., Leslie, D., Baddeley, R. & Farrell, S. Averaging versus sampling in collaborative judgement. *PsyArXiv,* 1–41 (2018).
16. Freund, Y., Mansour, Y. & Schapire, R. E. *Why averaging classifiers can protect against overfitting.* in *Eighth International Workshop on Artificial Intelligence and Statistics* (2001). `http://link.springer.com/10.1007/3-540-48219-9%7B%5C_%7D12%7B%5C%%7D0Apapers3://publication/doi/10.1007/3-540-48219-9%7B%5C_%7D12`.

17. Fugelsang, J. A. & Thompson, V. A. Strategy selection in causal reasoning: When beliefs and covariation collide. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* **54,** 15 (2000).

18. Gershman, S. J., Markman, A. B. & Otto, A. R. Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General* **143,** 182 (2014).

19. Gigerenzer, G. & Selten, R. *Bounded rationality: The adaptive toolbox* (MIT press, 2002).

20. Greenwald, A. G. & Banaji, M. R. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review* **102,** 4 (1995).

21. Hahn, U. & Chater, N. Similarity and rules: Distinct? Exhaustive? Empirically distinguishable? *Cognition* **65,** 197–230. ISSN: 00100277 (Jan. 1998).

22. Hahn, U. & Nakisa, R. C. German inflection: Single route or dual route? *Cognitive Psychology* **41,** 313–360 (2000).

23. Herzog, S. M. & von Helversen, B. Strategy Selection Versus Strategy Blending: A Predictive Perspective on Single- and Multi-Strategy Accounts in Multiple-Cue Estimation. *Journal of Behavioral Decision Making* **31,** 233–249. ISSN: 08943257 (Apr. 2016).

24. Herzog, S. M. & Hertwig, R. The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science* **20,** 231–237 (2009).

25. Hoffmann, J. A., von Helversen, B. & Rieskamp, J. Testing learning mechanisms of rule-based judgment. *Decision* **60,** 27–30. ISSN: 2325-9973 (Apr. 2019).

26. Hoffmann, J. A., von Helversen, B. & Rieskamp, J. Deliberation's blindsight: How cognitive load can improve judgments. *Psychological science* **24,** 869–879 (2013).

27. Hoffmann, J. A., von Helversen, B. & Rieskamp, J. Pillars of Judgment: How Memory Abilities Affect Performance in Rule-Based and Exemplar-Based Judgments. *Journal of Experimental Psychology: General* **143,** 2242–2261 (2014).

28. Hoffmann, J. A., von Helversen, B., Rieskamp, J., Weilbächer, R. A. & Rieskamp, J. Similar task features shape judgment and categorization processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **42,** 1193–1217. ISSN: 1939-1285 (June 2016).

29. Jacobs, R. A., Jordan, M. I. & Barto, A. G. Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks. *Cognitive science* **15,** 219–250 (1991).

30. Jacobs, R. A., Jordan, M. I., Nowlan, S. J. & Hinton, G. E. Adaptive Mixtures of Local Experts. *Neural Computation* **3,** 79–87. ISSN: 0899-7667 (1991).

31. Juslin, P., Karlsson, L. & Olsson, H. Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition* **106,** 259–298. ISSN: 00100277 (Jan. 2008).

32. Juslin, P., Olsson, H. & Olsson, A.-C. Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General* **132,** 133–156. ISSN: 0096-3445 (2003).

33. Kahneman, D. *Thinking, fast and slow* (Macmillan, 2011).

34. Kalish, M. L., Lewandowsky, S. & Kruschke, J. K. Population of Linear Experts: Knowledge Partitioning and Function Learning. *Psychological Review* **111,** 1072–1099 (2004).

35. Karelaia, N. & Hogarth, R. M. Determinants of linear judgment: a meta-analysis of lens model studies. *Psychological Bulletin* **134,** 404–426. ISSN: 0033-2909 (2008).

36. Karlsson, L., Juslin, P. & Olsson, H. Adaptive changes between cue abstraction and exemplar memory in a multiple-cue judgment task with continuous cues. *Psychonomic Bulletin and Review* **14,** 1140–1146. ISSN: 10699384 (2007).

37. Kruschke, J. K. ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review* **99,** 22–44. ISSN: 0033-295X (1992).

38. Kurvers, R. H. J. M. *et al.* Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences* **113,** 8777–8782. ISSN: 0027-8424 (Aug. 2016).

39. Larrick, R. P. & Soll, J. B. Intuitions about Combining Opinions: Misappreciation of the Averaging Principle. *Management Science* **52,** 111–127. ISSN: 00251909, 15265501 (2006).

40. Liaw, A., Wiener, M., *et al.* Classification and regression by randomForest. *R news* **2,** 18–22 (2002).

41. Liberman, V., Minson, J. A., Bryan, C. J. & Ross, L. Naive realism and capturing the "wisdom of dyads". *Journal of Experimental Social Psychology* **48,** 507–512 (2012).

42. Lieder, F. & Griffiths, T. L. Strategy selection as rational metareasoning. *Psychological Review* **124,** 762 (2017).

43. MacLeod, C. M., Hunt, E. B. & Mathews, N. N. Individual differences in the verification of sentence—picture relationships. *Journal of verbal learning and verbal behavior* **17,** 493–507 (1978).

44. Marewski, J. N. & Link, D. Strategy selection: An introduction to the modeling challenge. *Wiley Interdisciplinary Reviews: Cognitive Science* **5,** 39–59 (2014).

45. Marewski, J. N. & Schooler, L. J. Cognitive niches: an ecological model of strategy selection. *Psychological review* **118,** 393 (2011).

46. Mata, R., Schooler, L. J. & Rieskamp, J. The aging decision maker: cognitive aging and the adaptive selection of decision strategies. *Psychology and aging* **22,** 796 (2007).

47. Mellers, B. *et al.* Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science* **10,** 267–281 (2015).

48. Mellers, B. *et al.* Psychological Strategies for Winning a Geopolitical Forecasting Tournament. *Psychological Science* **25,** 1106–1115. ISSN: 0956-7976 (May 2014).

49. Minson, J. A., Liberman, V. & Ross, L. Two to tango: Effects of collaboration and disagreement on dyadic judgment. *Personality and Social Psychology Bulletin* **37,** 1325–1338 (2011).

50. Minson, J. A. & Mueller, J. S. The cost of collaboration: Why joint decision making exacerbates rejection of outside information. *Psychological Science* **23,** 219–224 (2012).

51. Nosofsky, R. M., Palmeri, T. J. & McKinley, S. C. Rule-plus-exception model of classification learning. *Psychological review* **101,** 53 (1994).

52. Pachur, T. & Olsson, H. Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology* **65,** 207–240 (2012).

53. Payne, J. W., Bettman, J. R. & Johnson, E. J. *The Adaptive Decision Maker* doi:10.1017/CBO9781139173933 (Cambridge University Press, 1993).

54. Pinker, S. *How the mind works* (Penguin UK, 2003).

55. Prelec, D., Seung, H. S. & McCoy, J. A solution to the single-question crowd wisdom problem. *Nature* **541,** 532–535. ISSN: 14764687 (2017).

56. Rieskamp, J. & Otto, P. E. SSL: a theory of how people learn to select strategies. *Journal of Experimental Psychology: General* **135,** 207 (2006).

57. Roberts, M. J., Gilmore, D. J. & Wood, D. J. Individual differences and strategy selection in reasoning. *British Journal of Psychology* **88,** 473–492 (1997).

58. Rodrigues, P. M. & Murre, J. M. Rules-plus-exception tasks: A problem for exemplar models? *Psychonomic Bulletin & Review* **14,** 640–646 (2007).

59. Rumelhart, D. E. & McClelland, J. L. On learning the past tenses of English verbs (1986).

60. Söllner, A., Bröder, A., Glöckner, A. & Betsch, T. Single-process versus multiple-strategy models of decision making: Evidence from an information intrusion paradigm. *Acta Psychologica* **146,** 84–96 (2014).

61. Speekenbrink, M. & Shanks, D. R. Learning in a changing environment. *Journal of Experimental Psychology: General* **139,** 266 (2010).

62. Sun, R., Merrill, E. & Peterson, T. From implicit skills to explicit knowledge: A bottom-up model of skill learning. *Cognitive science* **25,** 203–244 (2001).

63. Sun, R., Slusarz, P. & Terry, C. The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological review* **112,** 159 (2005).

64. Tetlock, P. E., Mellers, B. A. & Scoblic, J. P. Bringing probability judgments into policy debates via forecasting tournaments. *Science* **355,** 481–483 (2017).

65. Touron, D. R. & Hertzog, C. Distinguishing age differences in knowledge, strategy use, and confidence during strategic skill acquisition. *Psychology and aging* **19,** 452 (2004).

66. Von Helversen, B., Herzog, S. M. & Rieskamp, J. Haunted by a Doppelgänger: irrelevant facial similarity affects rule-based judgments. *Experimental Psychology* **61,** 12–22. ISSN: 1618-3169 (Jan. 2014).

67. Vul, E. & Pashler, H. Measuring the Crowd Within. *Psychological Science* **19,** 645–647. ISSN: 0956-7976 (July 2008).

A BASICS: A generalized mixture-of-experts model for strategy selection

In mixture-of-experts models, the probability P_c of each solution or output category c is defined as the summed probability $P_{c,s}$ of choosing this solution across all single experts or strategies s, weighted by the strength or activation a_s of each strategy, $P_c = \sum_s a_s * P_{c,s}$.

BASICS, **B**lending **A**nd **S**hifting **I**n **C**oordinating **S**trategies, assumes that the probability $P_{c,s}$ of observing solution c under strategy s is a function f_s of the problem activation x itself, defining how strongly different features or dimensions are activated, and a weight matrix $w_{c,x}$ defining how important different dimensions are according to this strategy: $P_{c,s} = f_s(w_{c,x}, x)$. Thus, BASICS is not limited to a particular set of competing strategies.

How likely each strategy is activated, the strategy activation a_s, depends upon a general preference β_s for the strategy s as well as contextual knowledge $\sum_h a_h * w_{s,h}$.

$$a_s = \frac{e^{\gamma * \sum_h a_h * w_{s,h} + \beta_s}}{\sum_s e^{\gamma * \sum_h a_h * w_{s,h} + \beta_s}} \tag{2}$$

In principle, this activated context, represented by the context activation a_h for context h, can reflect a variety of previously learned activation rules elicited by the current situation or task context and can thus be broadly defined as a function of the current problem activation x, $a_h = g(w_{h,x}, x)$. In the simplest case, the decision maker may remember an object as an exception, but may have also learned more complex contextual dependencies, such as "All objects bigger than X". Each context may then activate one or many associated strategies with the context weights $w_{s,h}$ storing the associations between each context and a strategy. For instance, the decision maker may apply a different strategy to all objects stored as exceptions (e.g. "Only objects looking like X"), or judge "All objects bigger than X" according to a different strategy than "All objects

smaller than X". How strongly the solution proposed by each strategy then influences the final solution depends upon its strategy activation relative to all other strategies, as implemented in the Softmax choice function which implies that all relative strategy activations sum up to 1. The strategy selectivity γ modulates the degree to which contextual knowledge preferably activates one strategy and ignores global strategy preferences or —on the contrary —the contextual knowledge is ignored and global strategy preferences dominate, eventually causing strategy-shifting or strategy-blending (see subsection A.2).

A.1 Adjusting the strategies' importance in response to feedback

If multiple strategies can be applied to one problem but only a few yield adequate problem solutions, it is essential to learn which strategy is the most appropriate one. BASICS learns to identify which strategy best solves the problem at hand by minimizing in each trial the summed error E across all strategies E_s, weighted by their strategy activation a_s, $E = \sum_s a_s E_s$. Each strategy commits a higher error if the strategy predicts the correct solution with a lower probability, $E_s = -\sum_c t_c * \ln P_{c,s}$, where $\ln P_{c,s}$ is the probability of predicting solution c for strategy s and t_c is the teacher in each trial. The teacher takes on the value 1 for the correct solution, 0 otherwise.

This feedback is used to learn across trials which strategy fares best at predicting the outcome by updating the general preference β_s for a strategy and the context weights $w_{s,h}$, the associations between each context and a strategy, in the opposite direction of the error using gradient descent.

$$\frac{\Delta E}{\Delta \beta_s} = -\left[\frac{\Delta a_1 E_1}{\Delta \beta_s} + ... + \frac{\Delta a_s E_s}{\Delta \beta_s}\right] \qquad (3)$$

The general strategy preference for each strategy β_s is adjusted more, if the strategy initially contributed to a higher extant to the final solution without being majorly responsible for the error on this trial, $\Delta\beta_s = \lambda_s a_s (E - E_s)$ with λ_s denoting the strategy learning rate. The context weights are adjusted in a similar fashion, but the magnitude of the update is larger within the activated context a_h, $\Delta w_{s,h} = \lambda_s \gamma a_h a_s (E - E_s)$. In addition to learning strategy preferences, BASICS learns and adjusts each single strategy across trials by updating how much weight each strategy assigns to different problem dimensions or any combination thereof.

A.2 Blending or shifting? The implications of varying strategy selectivity

Within BASICS, the strategy selectivity γ alters to what degree people predominantly blend the output of all strategies or shift between strategies depending upon the context. If strategy-blending is defined as integrating the solutions of all strategies, but ignoring any context-dependent knowledge, then a strategy selectivity approaching 0 introduces strategy-blending within each trial (cf. Equation 1): $\lim_{\gamma \to 0} a_s = e^{\beta_s} / \sum_s e^{\beta_s}$.

In each trial, the strategies are activated relative to the person's preference for this strategy without considering contextual knowledge. In a single trial, the probability of choosing a solution is then the average probability of choosing this solution across all strategies, weighted by the strategies' activation. Learning from feedback finally updates the general preference for each strategy β_s without updating any context-dependent knowledge in the form of context weights $w_{s,h}$.

At the opposite end of this spectrum, strategy-shifting emphasizes that people follow only a single strategy in each trial, but shift depending upon contextual knowledge. This strategy-shifting can be implemented by assuming an infinitely high strategy selectivity, $\lim_{\gamma \to \infty} a_s$. This implies that any general strategy preference does not influence strategy activation and only the context activation determines strategy choice, $\sum_h a_h * w_{s,h}$. In particular, only the strategy is activated, $a_s = 1$, that shares the strongest association with the current context, $max_s\{\sum_k a_h * w_{s,h}\}$. In each trial, the probability of choosing one solution

thus depends how likely each solution is chosen by the activated strategy. Yet, this high strategy selectivity has consequences beyond strategy activation in a single trial and affects how individuals learn from feedback. Importantly, if only one strategy is activated in this trial, the total error E reflects only a single strategy and is thus equivalent to the strategy-specific error E_s. Learning from feedback thus does no longer allow to attribute the error to different strategies, making it impossible to learn which strategy to use in a particular context in strategy-shifting.

B BASICS applied to human judgment

To model human judgment, BASICS assumes that people base their judgment at any point in time on two strategies: a rule-based and a memory-based strategy. Over trials, BASICS develops a preference for a memory-based over a rule-based strategy, thereby distinguishing between global and contextual strategy preferences.

B.1 Rule-based judgment

Rule-based strategies assume that people consider different aspects, or cues, that could influence their judgment, weight these cues by their importance, and sum up the weighted cue values. This idea has been formalized by portraying a rule-based judgment $\hat{j}_{\mathrm{RL}}$ for the object p as a linear, additive function of the cue values $x_{k,p}$ (with cue dimensions $1, ..., k$) weighted by their importance w_k, which can be mathematically modeled by a linear regression.

$$\hat{j}_{\mathrm{RL}} = \sum_k w_k \cdot x_{k,p} \tag{4}$$

with $x_{*,p}=\begin{bmatrix} x_{1,p} & ... & x_{l,p} & 1 \end{bmatrix}$ where l denotes the number of cues and 1 denotes the constant intercept. Here, we propose a capacity-constrained rule-based learner by imposing a capacity restriction, $\sum_k |w_k| < r$, on the cue weights [see 25, for a rationale of the restrictions]. The rule-based judgment $\hat{j}_{\mathrm{RL}}$ finally activates the discrete response category c, $a_{c,\mathrm{RL}} = -\frac{1}{2\sigma}(j_c - \hat{j}_{\mathrm{RL}})^2$. Response activations, $a_{c,\mathrm{RL}}$, are converted into response probabilities $P_{c,\mathrm{RL}}$ using the Softmax function.

B.2 Memory-based judgment

Memory-based judgment strategies assume that people compare the object under evaluation, the probe p, to all previously encountered objects stored in memory, the exemplars, and retrieve the judgments associated with the stored exemplars to estimate the final judgment $\hat{j}_{\mathrm{ML}}$. The more closely the probe's cue values $x_{k,p}$ (with cue dimensions $1, ..., k$) match the cue values of exemplar o $x_{k,o}$, the more strongly the corresponding exemplar is activated, a_o.

$$a_o = e^{-\rho \sum_k w_{\mathrm{ML},k} \cdot |x_{k,p} - x_{k,o}|} \tag{5}$$

A higher attention weight $w_{\mathrm{ML},k}$ for one cue k indicates that people pay more attention to that cue and, correspondingly, weigh this cue more heavily in the activation process. The memory sensitivity ρ finally modulates how specifically people activate single exemplars. Exemplars then activate the discrete judgment responses c, $a_{\mathrm{ML},c}$ depending on their own activation and their previously learned association with the judgment value, the association weights $w_{o,c}$.

$$a_{\mathrm{ML},c} = \sum_j a_o \cdot w_{o,c} \tag{6}$$

Activated exemplars also serve as a retrieval cue for strategy preferences, that is, previously encountered exemplars represent the task context, $a_h = a_o$. The context weights, $w_{s,h}$, store item-specific preferences for exemplar retrieval. Depending on which exemplars are activated in memory, the context weights facilitate or inhibit memory-based judgments. Response activations, $a_{c,\mathrm{EL}}$, are converted into response probabilities $P_{c,\mathrm{EL}}$ using the Softmax function.

B.3 Learning rule- and memory based strategies

During learning, BASICS minimizes the log loss function, weighing the probabilities of the judgment response $P_{c,s}$ for each strategy s with the strategy ac-

tivation, a_s. The context weights are adjusted using back-propagation after each trial (see A). Learning the rule-based strategy requires to learn over trials which features should be considered as more relevant for the judgment at hand, but also which features should be completely ignored. Learning the memory-based strategy requires to learn over trials which exemplars are associated with a certain judgment response and to shift attention towards more important cues. Finally, LEARN updates as well the cue weights, the association weights and the attention weights from the rule-based and the memory-based strategy, respectively.

C Experimental procedures

Ethical approval was provided by the University of Konstanz where the experiments were conducted. All participants provided informed consent before beginning the study.

C.1 Experiment 1: High contextual knowledge

C.1.1 Participants

103 participants were recruited from the participant pool of the University of Konstanz (83 females, M_{Age} = 21.7, SD_{Age} = 4.6). Participants received an hourly fee (8.50 €) or course credit for their participation in the experiment. In addition, they could earn a performance-dependent bonus(M = 3.23 €, SD = 0.38 €).

C.1.2 Design and material

The cover story in the multiple-cue judgment task asked participants to judge how much power different robots can use on a scale from 0 to 50. Participants saw pictures of these robots that varied on four different quantitative cues, the power modules, with six quantitative cue values, the power slots. Power modules were attached to the arms and legs of the robot and varied in color (orange, green, violet, blue) and shape (cross, rectangle, circle, triangle). Position, color, and shape of each cue was determined randomly. Higher cue values were always associated with higher power slots for each cue (or module). These pictorial cues could be used to predict the criterion (the power level).

In the training phase of the rule-plus-exception task, the power level of most robots followed a linear, additive function of the cues, $y = 4x_1+3x_2+2x_3+x_4$. However, two of the robots were exceptions and possessed a criterion value distinct from the linear, additive function. These exceptions were repeated either 5 times (low frequency) or 25 times (high frequency) in training. Two rule-following items were repeated with the same frequency, whereas the two new items were never presented in training. We defined exceptions, rule-following items, and novel items as centers of one subarea within multi-dimensional space, thus defining a prototype, and determined them randomly for each participant. Prototypes at minimum 8 units apart in multi-dimensional space. All items within the city-block distance $d \leq 4$ were classified to this prototype and not shown in training. Random rule items were selected from the remaining items. A normally distributed random error was added to all judgment criteria (M = 0, SD = 1).

In test, participants judged all prototypes and distortions of the prototypes to test for distance effects (see Supplemental Information). Randomly selected rule-following items and the two new prototypes served as control items. All items were repeated in familiarity-based decisions. To avoid repeating the items too often, we constructed four matched subsets of distortions for each prototype that were subsequently paired with the exact prototype and distortions from subsets of another prototype. This design resulted in 120 familiarity-based decisions.

C.1.3 Procedure

The training phase consisted of 250 learning trials, divided into 10 learning blocks with 25 trials each. In each trial, participants saw one robot on the left side of the screen and had to choose its criterion value on a half-circular scale ranging from 0 to 50 on the right. Afterwards participants received feedback about their own answer, the correct outcome,

and the points they earned. Random rule-following items, repeated rule-following items, and exceptions were randomly interspersed.

After 250 learning trials, participants moved on to the test phase that consisted of two different tasks. In the judgment task, participants made judgments in the same fashion as in training, but did not receive any feedback about the correct outcome or their performance. In familiarity-based choice, participants were repeatedly asked which of two presented robots looked more familiar to them, that is, more similar to the robots presented in training. Half of the participants completed the judgment task first, the other half completed the familiarity-based choice first.

To motivate participants to achieve a high judgment accuracy, they could earn points in each judgment trial depending on how much their judgment j deviated from the correct criterion y:

$$\text{Points} = 20 - \frac{(j-y)^2}{7.625} \tag{7}$$

This function was truncated so that participants could win at most 20 points in each trial and could not lose any points. At the end of the experiment, the points earned were converted into €(2000 points = 1 €).

C.2 Experiment 2: Low contextual knowledge

C.2.1 Participants

102 participants were recruited at the University of Konstanz (79 females, $M_{\text{Age}} = 23.3$, $SD_{\text{Age}} = 5.1$, for 1 participant demographic information was missing). Participants received an hourly fee (8.50 €) or course credit for their participation in the experiment. In addition, they could earn a performance-dependent bonus($M = 2.98$ €, $SD = 0.44$ €).

C.2.2 Design, material, and procedure

Design, procedure and material remained unchanged in Experiment 2, except for the following adaptations. Participants did not judge not the same exceptions in training, but exception items highly similar to each other. For this reason, we never presented the exception, the center of each prototype, in training, but only presented distortions, items close to the center, and repeated each distortion only once during training (see Supplemental Information). Thus each single item should not elicit a high familiarity. To facilitate the judgment task, we used only three prototypes, one for exceptions, one for repeated rule-following, and one for novel items. Repeated rule-following and novel items were selected in the same manner as exceptions. Participants judged 10 exceptions (and rule-following items) in the low frequency condition and 50 in the high frequency condition during training. In the test phase, participants judged new distortions from each prototype (51 items), among them the center of the prototype repeated six times, and 51 random rule-following items. In total, this scheme resulted in 204 judgments and 102 familiarity-based paired comparisons.

Supplemental Information to "Coordinating several mental strategies requires integration: Evidence from human judgment"

Janina A. Hoffmann *[1,2], Rebecca Albrecht[3], and Bettina von Helversen[4,5]

[1]University of Bath
[2]University of Konstanz
[3]University of Basel
[4]University of Bremen
[5]University of Zürich

August 7, 2020

1 Training and validating BASICS in human judgment

In a judgment task, BASICS should be able to learn to weigh the contributions of a memory-based and a rule-based strategy while simultaneously acquiring knowledge about each strategy. Model training served the purpose of identifying learning parameters for both strategies with the prerequisite that a) BASICS acquires both strategies at the same speed in training and b) predominantly approaches classical rule-based tasks with a rule-based strategy after training, and c) predominantly approaches classical memory-based tasks with a memory-based strategy after training.

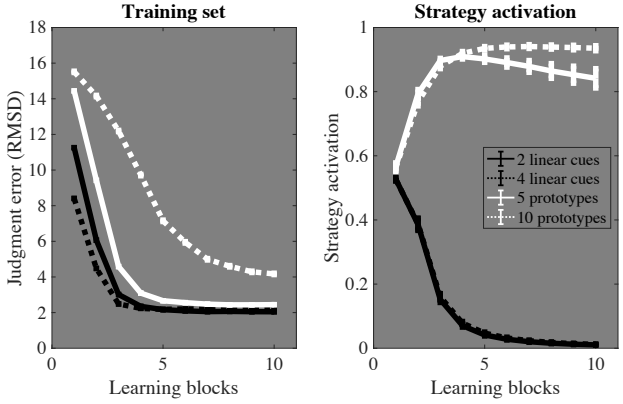
For this purpose, we trained BASICS simultaneously on four problems: two linear functions best approached by the rule-based strategy and two prototype tasks best approached by the memory-based strategy. We fixed the strategy learning rate λ_s to a low learning rate $\lambda_s = .05$ and the strategy selectivity γ to no preference for blending over shifting $\gamma = 1$. To estimate the parameter values for each strategy, we generated for each problem 1000 different training sets, each consisting of 250 randomly selected training objects with four cues

*[2]Janina A. Hoffmann, Department of Psychology, University of Bath, BA2 7AY, Bath, United Kingdom. E-mail: j.a.hoffmann@bath.ac.uk

taking on cue values from 0 to 5. In one linear environment, the judgment criterion was a linear function of two cues; in the second environment all four cues predicted the judgment criterion. Training objects and cue weights were randomly selected, but cue weights were restricted to the a scale from 0 to 50. In the prototype tasks, we selected 5 prototypes (or 10, respectively) that were strongly dissimilar from one another. The training objects were selected as random fluctuations from these prototypes adding normally distributed noise on each cue value ($M = 0$, $SD = .5$). The judgment criterion for each prototype was randomly assigned by dividing the scale into equally spaced intervals depending on the number of prototypes, assigning one prototype to each interval and finally selecting a random value from this interval for this prototype. A normally distributed noise was added to all judgment criteria ($M = 0$, $SD = .5$). Initial model validation was performed on 10000 newly generated validation sets, each consisting of the four initial problems. Using randomly drawn parameter sets from model training, we predicted judgment error and strategy activation on these new validation sets.

Figure 1 and **??** display how fast BASICS learns to solve each judgment problem (judgment error in the left panels) and to what degree it preferably activates the memory-based strategy (strategy activation in right panels) for the training sets (Figure 1) and the validation sets (Figure 1). Changes in judgment error (measured in RMSD in each training block) suggest that BASICS learns linear rules and 5 prototypes at a similar speed, but learns a task with 10 prototypes more slowly. Considering how strongly BASICS draws upon the memory- and the rule-based strategy in each problem suggests likewise that it successfully adapts to different judgment problems. Finally, BASICS learns to solve each problem in the validation sets with a slightly slower learning speed as in the original training sets, but with a sufficient level of accuracy. Likewise, it activates the same strategies in the new validation sets. In combination with its ability to replicate classical results in judgment research (see main text), these results indicate that BASICS provides a suitable tool to investigate strategy coordination in human judgment.

Figure 1: Learning performance in the training set for 250 training objects and strategy activation of the memory-based strategy.



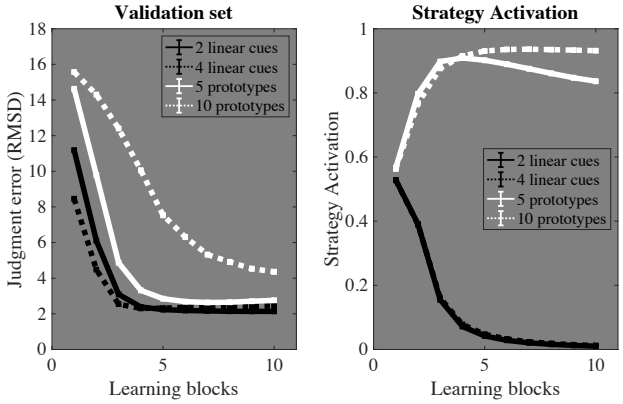


Figure 2: Learning performance in the validation set for 250 training objects and strategy activation of the memory-based strategy.

Initial simulations indicated, however, that varying strategy selectivity has little consequences for how individuals should learn to solve classical judgment tasks. BASICS learned to solve the four judgment problems [1] at the same speed and developed the same preferences for memory-based and rule-based strategies. Subsequent simulations identified rule-plus-exception tasks, tasks requiring applying a rule and simultaneously memorizing exceptions to this rule, as suitable to distinguish between varying levels of strategy selectivity. Figure 3 illustrates predictions for average judgment error and strategy preference in different rule-plus-exception tasks varying the frequency of exceptions belonging to one prototype during training (5 to 50 exceptions, different lines) and strategy selectivity (columns). At the end of training, strategy-blending (left graph) elicited a lower judgment error on rule-following items compared to items

belonging to the prototype and preference for the memory-based strategy only varied with the frequency of the exceptions, but not as a function of distance to the prototype. Strategy-shifting in all simulations elicited a high activation of the memory-based strategy (right graph). Selective activation (middle graphs) preferred the memory-based strategy more strongly for items more closely resembling the prototype, but predicted rule-based processing the more distant the item was from the prototype.

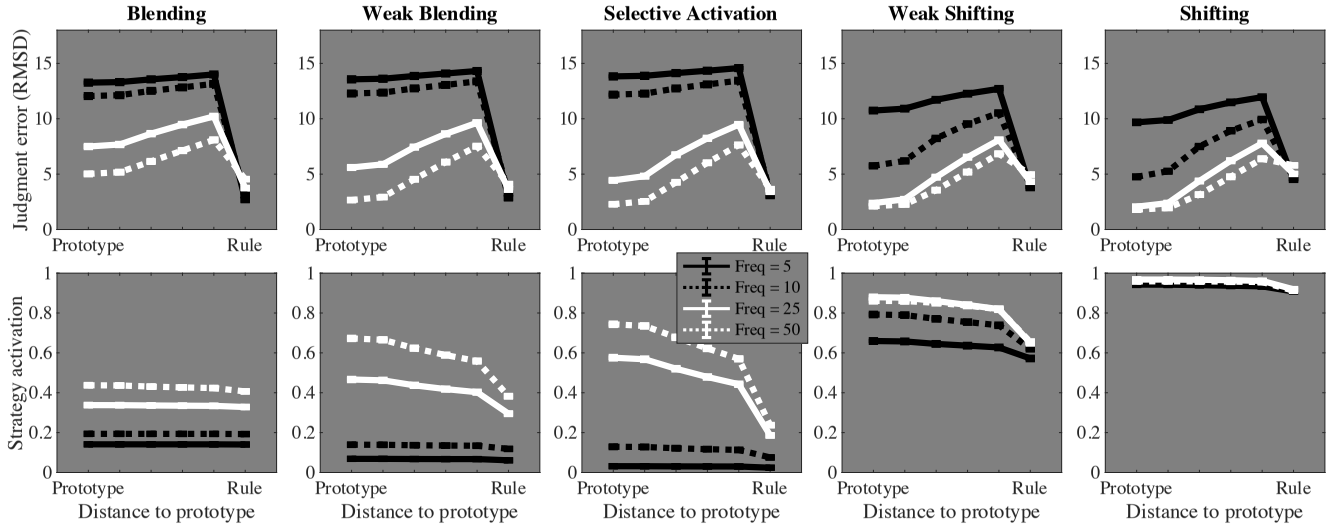


Figure 3: Judgment error (in RMSD) and strategy activation of the memory-based strategy for items belonging to the prototype and random rule-following items. Distance to the prototype is varied on the x-axis, frequency of exceptions during training in different lines. Different plots vary the strategy selectivity from Blending ($\gamma = .1$) to Weak Blending ($\gamma = .5$) to Selective Activation ($\gamma = 1$) to Weak Shifting ($\gamma = 5$) to Shifting ($\gamma = 10$).

2 Presented items as a function of distance

Table 1: Number of items selected in training and test depending on distance to the center of the prototype in Experiment 1.

Phase	Distance	Low frequency	High frequency
Training	0	5	25
Test	0	6	6
	1	4	4
	2	4	4
	3	8	8
	4	8	8
Familiarity	0	0	0
(matched subsets)	1	1	1
	2	1	1
	3	2	2
	4	2	2

2

Table 2: Number of items selected in training and test depending on distance to the center of the prototype in Experiment 2.

Phase	Distance	Low frequency	High frequency
Training	0	0	0
	1	1	3
	2	2	12
	3	3	15
	4	4	20
Test	0	6	6
	1	3	3
	2	12	12
	3	15	15
	4	15	15
Familiarity	0	2	2
(matched subsets)	1	1	1
	2	4	4
	3	5	5
	4	5	5

3 Additional result for experiment 2

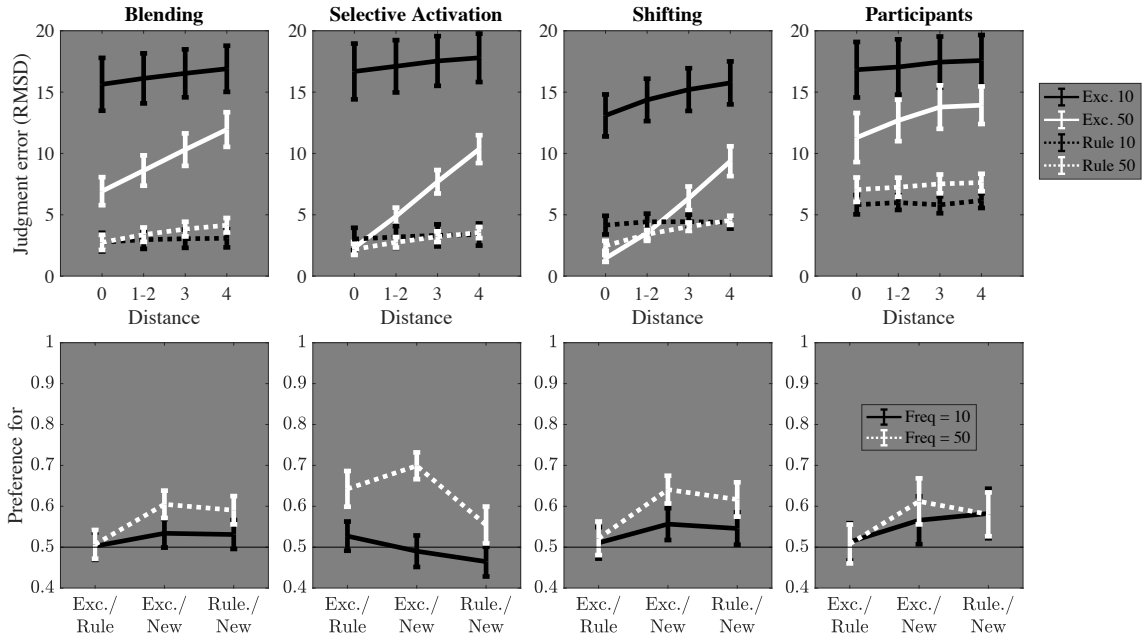


Figure 4: BASICS predictions for judgment error (upper panels) and familiarity-based choices (lower panels) as well as participant's judgment error (right upper panel) and participants' familiarity-based choices (right lower panel) for exceptions (Exc.), rule-following items (Rule), and new, unseen items (New). Participants (and the model) encountered 10 (black lines, or 50, white lines) similar exception items (and rule-following items) during training. Model predictions are shown for strategy-blending (left panel, $\gamma = 0.1$), selective activation ($\gamma = 1$), and strategy-shifting ($\gamma = 10$). In test, participants judged the prototype not presented in training (Distance = 0), but also items closely resembling the prototype (Distance = 1-4, distance 1 and 2 were collapsed to one data point). Judgment error was measured as the RMSD between the correct criterion in training and predicted judgments in test. Familiarity-based choices depict how often participants chose one item type (e.g. the exception item) as more familiar than another item type (e.g. rule following item. Error bars indicate 95 % confidence intervals.

References

1. Hoffmann, J. A., von Helversen, B., Rieskamp, J., Weilbächer, R. A. & Rieskamp, J. Similar task features shape judgment and categorization pro-

cesses. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **42,** 1193–1217. ISSN: 1939-1285 (June 2016).

Rebecca Albrecht-Dietsch

Education

University of Basel, Basel, Switzerland

PhD Student, Psychology, since September 2015.

- Topic: *Similarities in judgment and decision making*
- Supervision: Prof Dr. Jörg Rieskamp, Prof Dr. Bettina von Helversen,

University of Freiburg, Freiburg, Germany

PhD Student, Cognitive Science, January 2014 to August 2015.

- Supervision: PD Dr. Dr Marco Ragni

M.Sc., Computer Science, December 2013

- Thesis Topic: *A formal semantics for the cognitive architecture ACT-R*
- Supervision: Professor Andreas Podelski, Dr. Bernd Westphal
- Areas of Study: Artificial Intelligence and Software Engineering with an emphasis on modeling formalisms.
- Minor in Cognitive Science

B.Sc., Computer Science, March 2011

- Thesis Topic: *Modeling the Tower of London Task*
- Supervision: Professor Bernhard Nebel, PD Dr. Dr Marco Ragni
- Areas of Study: Software Engineering and Computer Engineering
- Minor in Cognitive Science

Projects

- **SNF Project: Modeling Human Judgment: Integrating Memory and Rule-based Processes**
  (research assistant) **September 2015 to August 2018**
- **SPP: New Frameworks of Rationality, Project: Shared Common Grounds of Qualitative and Quantitative Rational Reasoning**
  (research assistant) **January 2015 to July 2015**
- **SFB/TR 8, Project R8 – CSPACE**
  (research assistant) **February 2014 to December 2014**
- **SFB/TR 8, Project R8 – CSPACE**
  (student assistant) **March 2009 to September 2012**

Publications

[1] Rebecca Albrecht, Janina A. Hoffmann, Timothy J. Pleskac, Jörg Rieskamp, Bettina von Helversen. Competitive retrieval strategy causes multimodal response distributions in multiple-cue judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(6), 1064-1090. 2020.

[2] Rebecca Albrecht and Rul von Stuelpnagel. Memory for salient landmarks: Empirical findings and a cognitive model. *Spatial Cognition XI*, pages 1064–1090. 2018.

[3] Vincent Langenfeld, Bernd Westphal, Rebecca Albrecht, Andreas Podelski. But does it really do that? Using formal analysis to ensure desirable ACT-R model behaviour. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, pages 659–664. 2018.

[4] Rebecca Albrecht, Holger Schultheis, and Wai-Tat Fu. Visuo-Spatial Memory Processing and the Visual Impedance Effect. *Proceedings of the 37th annual meeting of the Cognitive Science Society*, pages 72–77. 2015.

[5] Matthias Frorath, Rebecca Albrecht, and Marco Ragni. Towards a unified reasoning theory: An evaluation of the Human Reasoning Module in Spatial Reasoning. *In Proceedings of the 13th International Conference on Cognitive Modeling (ICCM)*, pages 180 – 185. 2015.

[6] Rebecca Albrecht and Bernd Westphal. F-ACT-R: Defining the ACT-R Architectural Space. In *Supplement to Cognitive Processing 15*, pages 79 – 81. 2014.

[7] Rebecca Albrecht and Bernd Westphal. Analysing Psychological Theories with F-ACT-R. In *Supplement to Cognitive Processing 15*, pages 77 – 79. 2014. **Winner of the CSS Best Student Paper Award.**

[8] Felix Steffenhagen, Rebecca Albrecht, and Marco Ragni, M. Automatic Identification of Human Strategies by Cognitive Agents. In *KI 2014: Advances in Artificial Intelligence*, pages 62 – 67. Springer International Publishing. 2014.

[9] Rebecca Albrecht and Marco Ragni. Spatial Planning: An ACT-R model for the Tower of London Task. In *Spatial Cognition IX*, pages 222 - 236. Springer International Publishing. 2014.

[10] Rebecca Albrecht, Sven Brüssow, Christoph Kaller, and Marco Ragni. Using a Cognitive Model for an In-Depth Analysis of the Tower of London. In *Proceedings of the 33th annual meeting of the Cognitive Science Society*, pages 693–698. 2011.

Conference Abstracts

[11] Rebecca Albrecht and Steve Heinke. Expectation Formation of Decision Makers in the Economy. *Experimental Finance Conference 2018*, University of Heidelberg. 2018.

[12] Benedikt Solf, Rebecca Albrecht and Rul von Stülpnagel. Effects of Structurally and Visually Salient Landmarks on Memory for Turning Directions. *German Cognitive Science Conference*, University of Darmstadt. 2018.

[13] Vincent Langenfeld, Rebecca Albrecht, Bernd Westphal. Introducing Model Checking to Facilitate the Search for Spatial Representations in ACT-R. *German Cognitive Science Conference*, University of Darmstadt. 2018.

[14] Rebecca Albrecht, Janina A. Hoffmann, Timothy J. Pleskac, Jörg Rieskamp and Bettina von Helversen. Explaining quantitative judgments with a mixture model combining exemplar retrieval and cue-abstraction. *German Cognitive Science Conference*, University of Darmstadt. 2018.

[15] Rebecca Albrecht, Janina A. Hoffmann, Timothy J. Pleskac, Jörg Rieskamp and Bettina von Helversen. Explaining quantitative judgments with a mixture model combining exemplar retrieval and cue-abstraction. University of Marburg. 2018.

[16] Rebecca Albrecht, Bettina von Helversen, Janina A. Hoffmann, Timothy J. Pleskac, and Jörg Rieskamp. Explaining multiple cue judgment with a mixture model that combines exemplar with cue abstraction processes. *50th Annual Meeting of the Society for Mathematical Psychology.* University of Warwick, 2017.

[17] Janina A. Hoffmann, Rebecca Albrecht, and Bettina von Helversen, . Integrating cue abstraction with retrieval from memory: A learning approach. *50th Annual Meeting of the Society for Mathematical Psychology.* University of Warwick, 2017.

[18] Rebecca Albrecht, Bettina von Helversen, Janina A. Hoffmann, Timothy J. Pleskac, and Jörg Rieskamp. Explaining multiple cue judgment with a mixture model that combines exemplar with cue abstraction processes. *SPUDM.* University of Haifa, 2017.

[19] Janina A. Hoffmann, Rebecca Albrecht, and Bettina von Helversen, . Integrating cue abstraction with retrieval from memory: A learning approach. *SPUDM*. University of Haifa, 2017.

[20] Aljoscha Da Silva, Rebecca Albrecht, Matthias Frorath, Marco Ragni, and Lars Konieczny. Age-specific strategies in visuo-spatial planning. *Spatial Cognition*. Temple University, Philadelphia, 2016.

[21] Rebecca Albrecht, Bettina von Helversen, Janina A. Hoffmann, Timothy J. Pleskac, and Jörg Rieskamp. An exemplar-based random walk model for quantitative estimation. *58th Conference of Experimental Psychologists*. University of Heidelberg, 2016.

[22] Karoline Greger, Rebecca Albrecht, and Rul von Stülpnagel. Modeling route recall using landmarks. *58th Conference of Experimental Psychologists*. University of Heidelberg, 2016.

[23] Aljoscha Da Silva, Rebecca Albrecht, Marco Ragni, and Lars Konieczny. Altersabhängige Planungsstrategien für visuell-räumliche Probleme. *58th Conference of Experimental Psychologists*. University of Heidelberg, 2016.

[24] Rebecca Albrecht, Gießwein, Michael, and Bernd Westphal. Towards Formally Founded ACT-R Simulation and Analysis. *German Cognitive Science Conference*, University of Tübingen, 2014.

[25] Rebecca Albrecht, Felix Steffenhagen, and Marco Ragni. Identifying Inter-Individual Planning Strategies. *German Cognitive Science Conference*, University of Tübingen, 2014.

[26] Rebecca Albrecht and Marco Ragni. The Context Planning Effect: Memory Processing in Visuospatial Problem Solving. *Biannual Conference of the German Cognitive Science Society*. University of Bamberg, 2012.

Seminar talks, Conference Tutorials, and Organized Symposia

- Rebecca Albrecht and Mikhail Spektor. Cognitive modeling in Computer Science and Psychology: Bridging the gap. Organized Symposium at the German Cognitive Science Conference, University of Darmstadt, 2018.
- Explaining multiple cue judgment with a mixture model that combines exemplar with cue abstraction processes. Seminar talk at the University of Freiburg, Social Sciences and Methods, 2018.
- Explaining multiple cue judgment with a mixture model that combines exemplar with cue abstraction processes. Seminar talk at the University of Konstanz, Decision Sciences, 2017.
- Explaining multiple cue judgment with a mixture model that combines exemplar with cue abstraction processes. Seminar talk at the University of Zurich, Cognitive Psychology, 2016.
- Marco Ragni, Rebecca Albrecht, and Stefano Benatti. The Cognitive Architecture ACT-R with Cognitive Robotics as an Application. Tutorial held with the 35th German Conference on Artificial Intelligence. Saarbrücken, 2012.
- Marco Ragni and Rebecca Albrecht. The Cognitive Architecture ACT-R. Tutorial held with the Biannual Conference of the German Cognitive Science Society, Bamberg, 2012.

Teaching Assistance

University of Basel, Switzerland

- Experiment Programming (Theory Seminar, Autumn Term 2018 and Spring Term 2019)

University of Freiburg, Germany

- Computational Modeling in Cognition (Lecture, Winter Term 2014/2015)
- Formal Methods and Programming for Cognitive Scientists (Seminar, Winter Term 2014/2015)

Student Advising

University of Basel, Switzerland

- Mattias Brunner: The Credit Card Effect: Influence of Payment Mode on Purchasing Behaviour. A Literature Review with Perspectives on Virtual Goods. (B.Sc. Thesis, Psychology, University of Basel, 2019)
- Sarah Trefzer: Anwendung von psychologischen Konzepten in Videospielen. (B.Sc. Thesis, Psychology, University of Basel, 2019)

University of Freiburg, Germany

- Aljioscha DaSilva: The development of planning abilities in children under consideration of Gestalt and complexity. (M.Sc. Thesis, Cognitive Science, University of Freiburg, 2015)
- Matthias Frorath: Identification of Human Knowledge Representations in Planning Tasks. (M.Sc. Thesis, Computer Science, University of Freiburg, 2015)
- Karoline Greger: A Cognitive Model for Landmark Selection in Navigation Tasks. (Master Project, Cognitive Science, University of Freiburg, 2015)
- Aljioscha DaSilva: Belief Revision in school children. (Master Project, Cognitive Science, University of Freiburg, 2015)
- Matthias Häberle: A Comparison of Different Knowledge Representations in the Tower of London Task. (Master Project, Cognitive Science, University of Freiburg, 2015)
- Matthias Frorath: An Evaluation of the Human Reasoning Module. (Master Teamproject, Computer Science, University of Freiburg, 2014)
- Michael Gießwein: Formalization and Implementation of the ACT-R Declarative Module. (B.Sc. Thesis, Computer Science, University of Freiburg, 2014).

Tutoring

University of Freiburg, Freiburg, Germany
Responsible for tutoring, grading, exam preparation, course work preparation.

- Unified Modeling with UML — Winter Term 2013
- Empirical Methods in Psycholinguistics — Summer Term 2013
- Introduction into Cognitive Science 2: Language — Winter Term 2012
- Connectionist Cognitive Modeling — Winter Term 2012
- Symbolic Cognitive Modeling — Summer Term 2011, 2012, 2013
- Lisp Programming for Cognitive Scientists — Winter Term 2010, 2011
- Object-Oriented Programming with Java — Summer Term 2010
- Software Engineering — Summer Term 2010

- Theoretical Computer Science — Winter Term 2009
- Hardware Lab — Summer Term 2009
- Operating Systems — Winter Term 2008
- Computer Engineering — Winter Term 2008