Improving and Extending Models of Quantitative Judgments

David Izydorczyk

Inaugural Dissertation

Submitted in partial fulfillment of the requirements for the degree Doctor of Social Sciences in the DFG Research Training Group "Statistical Modeling in Psychology" at the University of Mannheim

Thesis Defense:
13.12.2022

Supervisors:
Prof. Dr. Arndt Bröder

Dean of the School of Social Sciences:
Prof. Dr. Michael Diehl

Evaluators:
Prof. Dr. Edgar Erdfelder
Prof. Dr. Daniel Heck

Examination Committee:
Prof. Dr. Arndt Bröder
Prof. Dr. Edgar Erdfelder
Prof. Dr. Daniel Heck

For my little family

Contents

Summary

How fast is this car approaching? What is the probability that it will rain today? How severe are the symptoms of this patient? Such quantitative judgments require inferring a continuous criterion from a number of cues or features of the judgment object (e.g., the color of the clouds). Judgments such as these are a central cognitive process which guides our decisions and behavior in our everyday life. For over half a century, researchers are investigating how people make such judgments, which information they rely on, how they combine different types of information, and how the environment or the task affect the processes underlying these judgments by using computational models of the theorized cognitive process.

It is the goal of my thesis to improve and extend these models of quantitative judgments. In three articles, I implement and test improved state-of-the art versions of existing models, highlight and solve issues in the way these models are currently used, and extend the scope and possibilities of these models of quantitative judgments. In the first manuscript, I develop, test, and apply a hierarchical Bayesian version of the RulEx-J model, which is used to measure the relative contribution of rule- and exemplar-based processes in people's judgments. The manuscript shows that the Bayesian RulEx-J model allows to estimate parameters more accurately and how it can be used to test hypotheses about latent parameters. The second manuscript shows that the current practice of not differentiating between direct retrieval of a trained exemplar and genuine judgments in the responses of participants leads to a biased estimation of parameters and reduced fit of exemplar-models. The manuscript also presents a solution to this problem by introducing a latent-mixture extended exemplar model which integrates a direct-recall process of trained exemplars. In the third manuscript, I demonstrate how to model people's judgments of even complex and realistic stimuli by extracting the necessary cues from pairwise similarity ratings.

In sum, the results of the three manuscripts described here contribute to the model-based study of the cognitive processes underlying people's judgments. By implementing state-of-the-art methods, improving upon current practices, and broadening the scope of the existing research, the results reported in this thesis add to the development, testing, and application of theories of quantitative judgments.

Manuscripts

This thesis is the result of research conducted in the context of the research training group "Statistical Modeling in Psychology" (SMiP) at the University of Mannheim. It is based on three articles, one of which has been published, one which is currently in press, and one which has been submitted for publication.

The three manuscripts aim at improving and extending models of quantitative judgments. First, I develop and test a hierarchical Bayesian implementation of the RulEx-J model to better measure the relative contribution of rule- and exemplar-based processes (Manuscript I). Second, I demonstrate how the analysis of a typical multiple-cue judgment experiment leads to biased parameter estimates of exemplar models (Manuscript II). Third, I show how models of quantitative judgments can be used to model judgments of non-artificial complex stimuli where the cues representing these stimuli and needed for the judgment models are not known beforehand (Manuscript III).

In the main text of this thesis, I provide a summary of the three manuscripts. Detailed description of the methods and statistical analyses can be found in the original manuscripts appended to this thesis.

MANUSCRIPT I

Izydorczyk, D., & Bröder, A. (in press). Measuring the mixture of rule-based and exemplar-based processes in judgment: A hierarchical Bayesian approach. *Decision.*

MANUSCRIPT II

Izydorczyk, D., & Bröder, A. (2021). Exemplar-based judgment or direct recall: On a problematic procedure for estimating parameters in exemplar models of quantitative judgment. *Psychonomic Bulletin & Review*, *28*, 1495–1513.

MANUSCRIPT III

Izydorczyk, D., & Bröder, A. (2022). *What is the airspeed velocity of an unladen swallow? Modeling numerical judgments of realistic stimuli.* [Manuscript submitted for publication]. Department of Psychology, University of Mannheim.

1 Introduction

In recent years, much of the discourse about problems of psychology as a science revolved about the "replication crisis", the low replicability rate of many well established findings (Klein et al., 2018; Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012; Wiggins & Christopherson, 2019). Many suggestions to solve the crisis focused on changes in research practices, data collection, data analysis, and publication procedures (e.g., Asendorpf et al., 2013; Benjamin et al., 2018; Dienes, 2016; Nosek et al., 2012). In addition, building on arguments already made decades ago, many authors have argued that the lack of well-specified and strong theories in psychology has also contributed to the replication crisis, or might even be the cause of it (Fried, 2020; Lykken, 1991; Meehl, 1978, 1990a, 1990b; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019). To improve upon this "theory crisis", researchers have called for the use of computational models as a solution to build stronger theories to advance psychology as a science (Borsboom et al., 2021; Farrell & Lewandowsky, 2010; Guest & Martin, 2021; Haslbeck et al., 2021; Oberauer & Lewandowsky, 2019; Smaldino, 2017, 2020; Wills & Pothos, 2012). Computational models are formalized instances of scientific theories (Guest & Martin, 2021; Lamberts, 2005). Compared to vague verbal theories, a theory formalized through mathematical equations or in a computational programming language makes the underlying ideas and assumptions more explicit, allows to derive precise and testable predictions, and to compare models of different theoretical accounts of the same phenomenon (Batchelder et al., 2017; Farrell & Lewandowsky, 2018).

One research area where computational models are already common is the area of quantitative or numerical judgments[1]. Examples of quantitative judgments are judging the size of a tree, the grade of a bachelor thesis, or the severity of a patient's disease. Quantitative judgments such as these, of different importance and in different contexts, are fundamental to our everyday life. Questions about the structure of the information people use to make their judgments, how they integrate information, and how the task, the environment or individual difference affect the judgment process have generated a wealth of research (see Bröder & Hilbig, 2017;

[1] Since the quantitative judgments of participants in the experiments reported in this thesis are expressed on a numerical scale, I use both terms ("quantitative" and "numerical") interchangeably in this context.

Goldstein & Hogarth, 1997, for an overview). In order to answer these questions, different computational models of the theorized cognitive processes have been developed (Albrecht et al., 2020; B. Brehmer, 1994; Bröder et al., 2017; Juslin et al., 2003; von Helversen & Rieskamp, 2008).

However, although the focus in the discussion of computational models in light of the replication crisis is often about translating theories into a testable model, psychological science can only progress by considering all aspects of the research process: the models themselves, as formalized instances of scientific theories; the experimental paradigms and measurements used to produce data; the statistical techniques used to relate the models to the data, to compare competing models, as well as to test hypotheses on parameters of these models; and the substantive questions that we want to answer with our theories (Nilsson et al., 2011; Smaldino, 2019; Wilson & Collins, 2019).

Therefore, the goal of this dissertation is to improve and extend existing models of quantitative judgments and the way they are used on multiple dimensions. Specifically, I develop and thoroughly test a hierarchical Bayesian version of the RulEx-J model, which allows more accurate estimates of latent parameters than the original formulation (Manuscript I); I highlight and solve a severe problem that results from the combination of exemplar models and the experimental paradigm in multiple-cue judgments (Manuscript II); and I show how cognitive models of quantitative judgments can be used to study judgments of complex naturalistic stimuli, expanding the possibilities of the existing models (Manuscript III). With this, my thesis contributes at improving the way cognitive models are used in research on quantitative judgments and thereby paving the way for future theoretical developments and new possibilities of applications for models of quantitative judgments.

In Chapter 2, I first give a brief overview of the history of research on quantitative judgments. Next, I present the general experimental paradigm used in the research on quantitative judgments, which is also used in all three manuscripts. Further, I give a brief introduction into the different cognitive process people are assumed to use to make their judgments and show how they are modeled. Finally, I review which factors influence which processes people rely on, how both processes interact with each other, and how the relative contribution of each process can be measured through the RulEx-J model.

In Chapter 3, I present different ways to improve and extend models of quantitative judgments in three manuscripts. In Manuscript I, I develop and thoroughly test

a hierarchical Bayesian implementation of the RulEx-J model, which allows estimating parameters more accurately and enables to better test hypothesis of differences in parameters between groups. In Manuscript II, I show that the combination of experimental paradigm and exemplar model as used in the multiple-cue judgment literature leads to biased estimation as well as impaired validity of parameters. Finally, in Manuscript III, I extend the possibilities of the existing models of quantitative judgments by showing how they can be used to investigate the underlying processes of judgments of natural stimuli where the cues are not known beforehand.

In Chapter 4, I conclude this thesis by discussing the implications of all three manuscripts for the field of quantitative judgments and by addressing open questions and highlighting avenues for future research.

With this thesis, I uncover and solve several weaknesses in the way cognitive models are currently used in the field of quantitative judgments. I therefore not only advance the current state of the literature on quantitative judgments, but also contribute to establishing and testing theories about the processes underlying people's judgments. In addition, my thesis expands the possibilities of the field by opening up new research areas, where the existing theories and models can be applied.

2 Theoretical Foundations

The beginning of research on quantitative multiple-cue judgments can be traced back to one of the most influential theoretical contributions to the field of Judgment and Decision-Making (JDM) in general: Egon Brunswik's lens model (Brunswik, 1952; Goldstein & Hogarth, 1997). Brunswik (1952, 1955) proposed that objects in the environment and their properties, such as the size of an object, the intelligence of a person, or the calorie content of a dessert, cannot be perceived directly by our sensory organs. Rather, these distal variables have to be inferred on the basis of proximal cues, which a person can perceive directly (e.g., the size of the retinal image of an object, the vocabulary a person uses, the sweetness of the dessert). However, the cues in the environment are only probabilistically related to the distal variable, since each cue contains only incomplete information about it. Therefore, people can integrate multiple cues in order to construct a judgment of the distal variable (for an overview see Goldstein, 2004; Hammond and Stewart, 2001).

Hammond (1955) suggested that Brunswik's principles of perception could be used to study clinical judgments: A clinician cannot perceive the underlying condition of a patient directly, but has to rely on multiple and, in isolation, ambiguous cues to come to a diagnosis. By using linear regression models, Hammond (1955) investigated how accurately clinical psychologists are able to judge patient's IQ scores based on their Rorschach tests and what specific information the clinicians relied on to make these judgments. Hammond's usage of the Brunswikian approach to study more complex judgments generated a wealth of research and new methodological developments, which were later synthesized in the *Social Judgment Theory* (SJT, B. Brehmer & Joyce, 1988; Doherty & Kurz, 1996; Hammond et al., 1975).

The SJT assumes that people's judgments of a criterion (i.e., the quantity being estimated) are based on the integration of multiple sources of information (i.e., the cues) and that linear regression models and related approaches can be used to analyze the judgments of participants (Cooksey, 1996; Doherty & Kurz, 1996; Hammond et al., 1975). In the following years, SJT has been applied in a variety of different domains, such as teachers' grading policies (A. Brehmer, 1988), personality judgments (Hirschmüller et al., 2013), metamemory (Bröder & Undorf, 2019), employment interviews (Gorman et al., 1978), risk judgments (Earle & Cvetkovich,

1988), and meteorological forecasting (Stewart, 1989), to investigate how accurate people's judgments are, which cues they use to make their judgments, and how people combine information from different cues (for on overview, see Brehmer & Brehmer 1988, Doherty & Kurz, 1996, Karelaia & Hogarth, 2008).

Over the decades, the focus of judgment research shifted from investigating how accurate people's judgments are and what cues they use, to understanding the psychological processes behind these judgments, how the task or individual differences influence the judgment process, and how different judgment processes interact. In line with research on category learning, two qualitatively distinct types of processes have been proposed, which rely on different knowledge representations: *Rule-based processes* and *exemplar-based processes* (Allen & Brooks, 1991; Erickson & Kruschke, 1998; Hahn & Chater, 1998; Juslin et al., 2003; Karlsson et al., 2008; Sloman, 1996). This shift in research focus was also paralleled by changes in the experimental paradigm used to investigate these new research questions. These changes in the experimental paradigm, however, come along with certain problems, as I will show in Manuscript II and III of this thesis. In the next sections, I first briefly introduce the general research paradigm used in the current multiple-cue judgment literature and thus the experiments reported in this thesis. Next, I introduce rule- and exemplar-based processes and give an overview of research investigating how the environment and task structure influence which process is used by participants and how these processes interact. Finally, I introduce the RulEx-J model which is one of the most recent models of quantitative judgments and part of Manuscript I and III of this thesis.

2.1 The Multiple-Cue Judgment Experiment

Brunswik's work and the research around the SJT are also the foundation for the experimental study of quantitative multiple-cue judgments, where participants rely on multiple cues to make their judgments on a continuous scale. While this traditional research on multiple-cue judgments has typically involved probabilism (i.e., cues are only probabilistically related to the criterion) and a large set of unique and complex stimuli with many cues (e.g., Einhorn et al., 1979; Hoffman, 1960), the current research in the last two decades after the work of Juslin et al. (2003) orients itself towards the experiments conducted in categorization research by using few simple artificial stimuli with a low number of deterministic cues (e.g., Estes, 1994;

Hoffmann et al., 2014; Juslin et al., 2003; Medin & Schaffer, 1978). Table 1 shows exemplary stimuli which are similar in their complexity and structure to the stimuli used in many multiple-cue judgment experiments (e.g., Bröder et al., 2010; Hoffmann et al., 2018; Trippas & Pachur, 2019). In a multiple-cue judgment task, people are asked to judge the criterion value of several stimuli, which can differ on multiple cues. For instance, the fictitious flowers in Table 1 can differ on three binary features, petal color (red/blue), leave form (thin/thick), and petal form (round/star-shaped). The criterion (e.g., the price of the flower) and the cues are deterministically related through the linear additive rule $c = 1 + 2 \times \text{cue}_1 + 3 \times \text{cue}_2 + 4 \times \text{cue}_3$.

Table 1

Example stimuli of an multiple-cue judgment experiment

Stim.	Cue 1 leave form	Cue 2 petal form	Cue 3 petal color	Crit. price in €	Training
	0	0	0	1	
	1	0	0	3	✓
	0	1	0	4	✓
	1	1	0	6	
	0	0	1	5	✓
	1	0	1	7	
	0	1	1	8	✓
	1	1	1	10	

Note. The criterion and the cues are related through the function $c = 1 + 2 \times \text{cue}_1 + 3 \times \text{cue}_2 + 4 \times \text{cue}_3$.

The typical experiment consists of two phases, a training phase and a testing phase. In the training phase of the experiment, participants are presented with a number of training stimuli (the exemplars). For example, in Table 1 four of the eight possible stimuli are selected as exemplars. In order to learn the exemplars, their criterion values, and the relationship between cues and criterion values, participants repeatedly have to judge the criterion value of these exemplars (i.e., the price in

this example) over the course of multiple blocks and they receive feedback about the correct criterion values. In the testing phase, participants then have to judge old stimuli (the exemplars) and new stimuli.

2.2 Rule-Based Processes

In line with the SJT, rule-based process models assume that people combine and integrate cue information according to some abstracted rule to make a judgment (Juslin et al., 2003). The rule, according to which the information of multiple cues is integrated, is often assumed to be a linear additive function (Einhorn et al., 1979; Hoffmann et al., 2019; Juslin et al., 2008; Juslin et al., 2003). For instance, if people are asked to judge the calorie content of a dessert they might combine the cues "sweetness", "amount of cream", and "fruits" in a linear addition fashion. The two main reasons for this often assumed linear additive rule are the limited processing capacity of the cognitive system and the (experimental) environment itself.

According to Juslin et al. (2008), capacity limitations of our cognitive system constrain the judgment process to a serial and additive integration of cues. This view is supported by research showing that people are quite good at learning linear additive rules and that they show a preference for using these types of rules (e.g., Ashby et al., 2001; B. Brehmer, 1974, 1994; Fischbein et al., 1985; Hoffmann et al., 2014; Hoffmann et al., 2018; Kalish et al., 2004; Mcdaniel & Busemeyer, 2005). Further, people's explicit judgment rules or cue weights are often compatible with linear regression models (Einhorn et al., 1979; Lagnado et al., 2006). In addition, although people are in principle able to learn non-linear, multiplicative or more complex rules, they have more difficulties doing so (Bott & Heit, 2004; B. Brehmer, 1969; Hammond & Summers, 1965; Mellers, 1980). Finally, the actual relationship between cues and criterion values is often linear in the everyday environment (B. Brehmer, 1994) and even more so in experimental settings (e.g., Bröder & Gräf, 2018; Juslin et al., 2003; von Helversen, Herzog, et al., 2014; Wirebring et al., 2018).

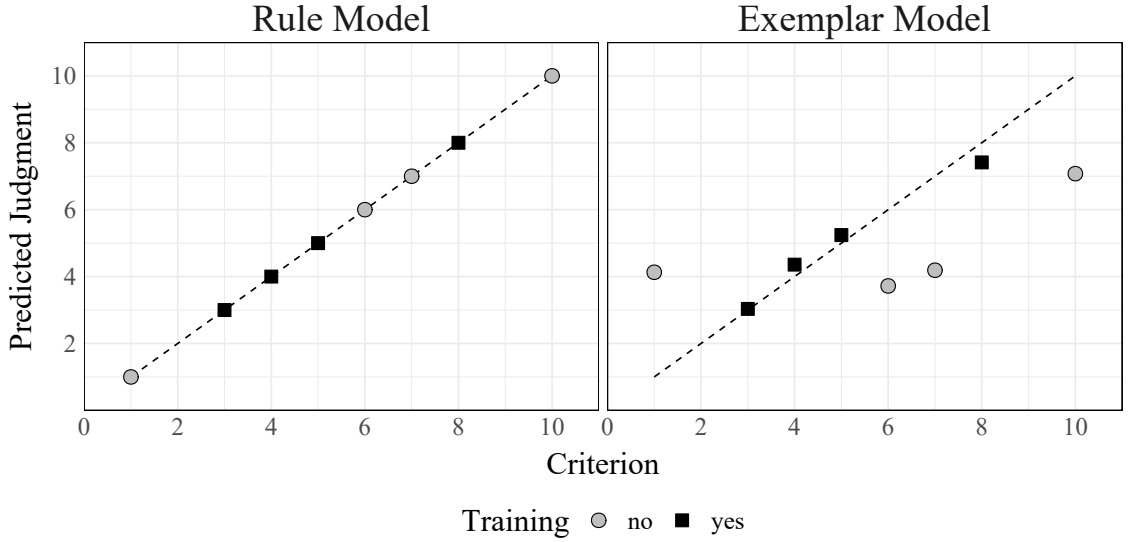
Juslin et al. (2003) formalized the assumption of an additive linear integration of multiple cues as follows:

$$J_p = w_0 + \sum_{i=1}^{n} \text{cue}_i \times w_i, \tag{1}$$

where J_p is the judged criterion of an object p (the probe) based on the intercept w_0 and the cue weights w_i corresponding to the n cues. This rule-based model, sometimes referred to as *cue abstraction model*, is quite flexible and does not necessarily imply a compensatory processing of all cues, but can also mimic simpler strategies or heuristics (like the lexicographic rule) focusing on one or only few cues by choosing appropriate (zero) cue weights (Bröder, 2000; Gigerenzer & Goldstein, 1996; Gigerenzer et al., 1999). The predictions of the rule-based model in Equation (1) for the exemplary experiment in Table 1 are shown in Figure 1. Since the training stimuli in Table 1 allow a perfect abstraction of the linear rule (i.e., perfect recovery of the cue weights), the predictions of the rule-based model are identical to the actual criterion values.

Figure 1

Predictions of the rule- and exemplar-based models based on the training stimuli in Table 1



Note. Predictions for the exemplar model are made with an assumed $s = .10$.

In general, it is often assumed that rule-based processes (especially learning and execution) pose high working memory demands (e.g., Ashby & O'Brien, 2005; B. Brehmer, 1994; Hoffmann et al., 2019; Juslin et al., 2008). One reason for that is that it has been suggested that people learn bivariate cue-criterion relationships (i.e. the cue weights w_i) by comparing sequentially presented objects and relating the difference in the criterion values of the objects to the difference in cue values, resulting in a trial-by-trial updating of cue weights during learning (B. Brehmer,

1974; Hoffmann et al., 2019; Juslin et al., 2008). Therefore, the previous judgment object(s) and their criterion values have to be maintained in working memory to then make a cue-wise comparison with the next judgment object. In addition, when the judgment is made, the cues have to be mentally integrated according to the learned rule. Recent empirical evidence supports the assumption that working memory is crucial in rule-based processes. For instance, Hoffmann et al. (2019) found that a formal learning model incorporating working memory constraints best predicted participants judgments in a task where the environment (i.e., the rule relating the cues to the criterion) changed after 200 trials and thus participants had to relearn the cue weights. In addition, several studies have shown that induced cognitive load or lower individual working memory capacity are related to lower judgment accuracy when people relied on a rule-based process (Hoffmann et al., 2014; Juslin et al., 2008; McDaniel et al., 2014).

Although rule-based processes are successful in predicting peoples judgments in many different tasks and situations, already Hammond and Brehmer (1973) proposed that in some tasks people rather rely on specific memories than on rules. Correspondingly, Juslin and colleagues (2008, 2003) suggested that judgments following more complex relationships of cues and criterion are actually not based on abstracted rules, but on past exemplars stored in memory. Drawing on categorization research, they provided a more formal account of this type of process in the form of exemplar models, which do not assume that people use abstracted rules to make their judgments, but instead that they rely on the storage and retrieval of past instances.

2.3 Exemplar-Based Processes

The assumption behind exemplar models is that people store previously encountered objects and their criterion values (i.e., the exemplars) as separate traces in episodic long-term memory (Estes, 1986; Hintzman, 1986; Medin & Schaffer, 1978). In line with Brunswik's assumption about the perception of distal variables, it is assumed that these stored exemplars are represented in memory by a number of attributes or cues (Bower, 1967; Estes, 1994; Fiedler, 1996). A new, to-be-judged object (i.e., the probe) acts as an retrieval cue to access the information stored in memory. The judgment of this new object is made by retrieving previously encountered exemplars from memory and then integrating the criterion values of these past exemplars based

on their similarity to the probe, where more similar exemplars have more impact on the judgment. Based on Shepard's law of generalization (Shepard, 1957, 1987), it is assumed that similarity is a function of the distance in psychological space between the probe and an exemplar based on the feature- and cue-dimensions which are used to mentally represent these objects (Medin & Schaffer, 1978; Nosofsky, 1986). Staying with the example of judging the calorie content of a dessert, when judging a new dessert, one would recall past desserts where the calorie content was (approximately) known (e.g., from the nutrition facts label). The calorie content of the new dessert will then be judged according to the similarity of this new dessert to the past desserts, whereby more similar past desserts will have a stronger influence on the judgment than dissimilar ones.

Although exemplar models originated and are most prominent in the research area of categorization (Medin & Schaffer, 1978; Nosofsky, 1986), they have been used in a variety of domains, such as memory (Hintzman, 1984), function learning (DeLosh et al., 1997), associative learning (Jamieson et al., 2012), social judgments (Fiedler, 1996; Smith & Zárate, 1992), decision-making (Bröder et al., 2010), Bayesian inference in human cognition (Shi et al., 2010), and language (Goldinger, 1996). Juslin and colleagues (2003, 2002) were the first to apply exemplar-models to quantitative judgments. They showed that exemplar-based models could predict people's judgment better than the hitherto common rule-based models when the relationship between cues and criterion was non-linear, which generated a wealth of research using exemplar-based models to study multiple-cue judgments (e.g., Hoffmann et al., 2013; Mata et al., 2012; Pachur & Olsson, 2012; Rosner & von Helversen, 2019; von Helversen & Rieskamp, 2008; Wirebring et al., 2018).

The formal exemplar model used by Juslin et al. (2003), which is also used in Manuscripts I and II of this thesis, is based on the *Context Model* of Medin and Schaffer (1978), but extended to account for continuous judgments (Juslin et al., 2003; Juslin & Persson, 2002). The context model is the first mathematical formulation of an exemplar-based model, which was later further developed into the Generalized Context Model (GCM, Nosofsky, 1984, 2011; Wills & Pothos, 2012). The context model as well as related models used in the multiple-cue judgment literature, implicitly assume an integrative retrieval of exemplars where all previously encountered exemplars and their criterion values are retrieved from memory and then integrated into the final judgment (cf., Albrecht et al., 2020; Nosofsky & Palmeri, 1997). The similarity of the probe to each of the exemplars acts as a

weight in the integration of all exemplar criterion values into the final judgment. More similar exemplars receive more weight and thus their criterion values have a higher impact on the final judgment (Estes, 1994; Medin & Schaffer, 1978). Formally, the similarity S between a probe p and an exemplar e is computed by the multiplicative similarity rule for binary cues (Medin & Schaffer, 1978):

$$S = \prod_{i=1}^{n} d_i \text{ with } d_i = \begin{cases} 1 \text{ if } p_i = e_i \\ s_i \text{ if } p_i \neq e_i \end{cases} \tag{2}$$

where n is the number of cues or features used to represent the objects. For each cue i, it is determined whether the cue values of the probe p and the exemplar e are equal. If they are equal, d_i has the value one, otherwise, d_i equals s_i. The s_i parameters are defined on the interval $[0, 1]$ and determine how strongly a mismatch of objects on the corresponding cue affects the overall similarity S. The closer s_i gets to zero, the more important the cue i becomes. Also, the closer s_i is to one, the more irrelevant the corresponding cue is. According to Medin and Schaffer (1978), the s_i parameters can be interpreted as attention parameters, where the s_i parameter of a cue i is lower when the dimension is attended to and thus the perceived similarity of two stimuli is lower, when the stimuli differ on this cue (but see Izydorczyk & Bröder, 2021, for a different interpretation). For instance, according to Equation (2), the similarity between the first two stimuli in Table 1 is $S = s_1 \times 1 \times 1 = s_1$, since only the first cue is different in both objects. When the s_1 parameter is low, indicating a high attention to this dimension, the similarity between probe and exemplar is also low. The original context model assumes that each cue has a different attention parameter (Medin & Schaffer, 1978). However, empirical evidence showed that a simplified version with a single s parameter (i.e., $s_i = s$) often provides a better or equal fit to actual judgment data than the more complex version with separate s parameters for each cue dimension (Hoffmann et al., 2013, 2014; von Helversen & Rieskamp, 2008, 2009a). Thus, the simplified version with only a single s parameter is used often instead of the full model. The judged criterion value of the probe J_p is then a similarity-weighted average of all k exemplars:

$$J_p = \frac{\sum_{j=1}^{k} S_j \times c_j}{\sum_{j=1}^{k} S_j} \tag{3}$$

where c_j is the criterion value of an exemplar j. Figure 1 shows the prediction

of the exemplar model for the example stimuli in Table 1, assuming a constant s of .1. There are two important differences when the predictions of the exemplar model in Figure 1 are compared with the predictions of the rule model. First, while the prediction accuracy of the rule model does not differ between trained exemplars (i.e., old items) and non-trained stimuli (i.e., new items), the prediction accuracy of the exemplar model is higher for old items than for new items. Second, the rule model has no difficulties extrapolating beyond the range of criterion values of the exemplars. However, the exemplar model is not able to extrapolate beyond the range of learned criterion values and thus predicts large judgment errors for the two most extreme stimuli (the flowers with the lowest and highest price) in Table 1 (DeLosh et al., 1997; Juslin et al., 2008).

Whereas rule-based processes of quantitative judgments strongly depend on working memory, exemplar-based processes are assumed to depend more on episodic memory, since they rely on the storage and retrieval of exemplars. This assumption is supported by empirical evidence showing that people with better episodic memory rely more on exemplar-based processes when making their judgments (Hoffmann et al., 2014). Also, several fMRI studies observed activity in brain regions associated with episodic memory when people relied on exemplars to make their judgments (Stillesjö et al., 2019; Wirebring et al., 2018). In addition, Hoffmann et al. (2018) found that a long retention interval of one week led to a greater decrease in judgment performance when people relied on an exemplar-based process compare to a rule-based process. That is presumably because people who rely on an exemplar-based process have to remember all (or at least some) learned exemplars, whereas people who relied on a rule-based process only have to remember the abstracted rule. Furthermore, the retrieval and similarity-weighted integration of exemplars when making judgments was demonstrated in an eye-tracking study by Rosner and von Helversen (2019). Using the looking-at-nothing paradigm (Richardson & Spivey, 2000), Rosner and von Helversen (2019) found that when judging the suitability of new fictitious job candidates participants looked more in the area of the screen were similar exemplars were presented during the training phase compared to less similar exemplars, indicating retrieval of these exemplars from memory.

2.4 Interaction of Rule- and Exemplar-Based Processes

Over the last decades, converging evidence from research on categorization, judgment, and decision making has shown that rule- and exemplar-based processes are indeed two qualitatively distinct processes which use different representations of knowledge and information, rely predominantly on different cognitive resources, and involve different brain regions (Allen and Brooks, 1991; Ashby et al., 1998; Hoffmann et al., 2014; Juslin et al., 2008; Pachur and Olsson, 2012; Sloman, 1996; von Helversen, Karlsson, Rasch, et al., 2014, but see Love et al., 2004; Schlegelmilch et al., 2021). Since the first introduction of exemplar-based models to the area of multiple-cue judgments (Juslin et al., 2003), most of the recent multiple-cue judgment research is now concerned with the questions of which factors influence whether a rule- or exemplar-based process is used in a given task or how people integrate the two types of processes. One might summarize the empirical results as following.

When people learn to judge objects where the relationship between the cues and the criterion is unknown, people try to abstract rules of how cues and the criterion are related by continuously updating the cue weights based on received feedback (Hoffmann et al., 2019). Simultaneously, memory traces of the encountered objects are stored in episodic memory (Hintzman, 1984). It is often assumed that people have a "rule-bias", that is, they initially try to rely on rule-based processes to make their judgments, but switch to exemplar-based processes based on task demands and their performance in the task (Juslin et al., 2008; Karlsson et al., 2008; Rieskamp & Otto, 2006). This is in line with an abundance of evidence showing that people tend to rely more on rule-based processes, when the environment, the cue format, learning task, or the provided feedback make it easier to abstract a rule (e.g., Juslin et al., 2003; Trippas & Pachur, 2019; von Helversen et al., 2013). For instance, several studies have found that when the relationship between cues and environment is linear (e.g., $c = 4\text{cue}_1 + 3\text{cue}_2 + 2\text{cue}_3 + 1\text{cue}_4$, Hoffmann et al., 2016), most people make their judgments according to a rule-based process. However, when the relationship between cues and environment is non-linear or multiplicative (e.g., $c = \frac{1}{8.5} \times [4\text{cue}_1 + 3\text{cue}_2 + 2\text{cue}_3 + 2\text{cue}_1\text{cue}_2\text{cue}_3 + \text{cue}_2\text{cue}_3\text{cue}_4]$, Hoffmann et al., 2016), people rely on exemplar-based processes (Hoffmann et al., 2014, 2016; Juslin et al., 2008; Karlsson et al., 2007; Olsson et al., 2006). In addition, giving people information about the direction and importance of cues also fosters the reliance on

rule-based processes (Bröder et al., 2010; Platzer & Bröder, 2013; von Helversen et al., 2013; von Helversen & Rieskamp, 2009a).

For exemplar-based processes, it seems important that the task, the environment, and the stimuli allow participants to form strong representations of specific exemplars, as well as an efficient retrieval of exemplars. For instance, having more distinguishable exemplars, a lower number of exemplars during training, a longer training phase, or more experience have been shown to promote reliance on exemplar-based processing, presumably because these factors increase the strength and discriminability of individual exemplars in memory (Johansen & Palmeri, 2002; Rouder & Ratcliff, 2004; Thibaut et al., 2018). In addition, increased cognitive load also decreases the accuracy of exemplar-based judgments (Hoffmann et al., 2014; Juslin et al., 2008), potentially by hindering the retrieval of past exemplars from memory or their integration (Anderson et al., 1996; Unsworth et al., 2013).

Despite these task or environmental factors influencing the main mode of processing, individual differences also seem to play a role. For example, most studies find large individual variability in what process people rely on in a given condition of the experiment (e.g., Hoffmann et al., 2016; Izydorczyk & Bröder, in press; Nosofsky & Hu, 2022; Rouder & Ratcliff, 2004). Also, whereas most adults seem to adapt their judgment strategy based on the task demands, younger children (9-11 years) rely more on exemplar processing compared to adults (von Helversen et al., 2010), whereas older adults ($>$ 60 years) stick to rule-based processing (Mata et al., 2012), which might be related to differences in working memory and episodic memory capacity (Hoffmann et al., 2016). In addition, McDaniel et al. (2014) found that people who were identified as rule-learners in a function learning task also relied more on rule-based processes in a subsequent categorization task, indicating differences in individual preferences (cf. Hoffmann et al., 2016).

2.4.1 Integration of Rule- and Exemplar-Based Processes

Initially, it was (implicitly) assumed that there is a "division of labor" between rule- and exemplar-based processes, where individuals would select one process based on task demands or other factors and then rely only on this process to make their judgments (Juslin et al., 2008; Juslin et al., 2003; Karlsson et al., 2008). The methods used to determine whether participants relied on rule- or exemplar-based processing in a given task reflected this dichotomization of the judgment process. For example,

researchers would classify participants as users of a rule- or exemplar-based strategy (reflecting the corresponding cognitive process) based on the best-fitting model (e.g., Bröder et al., 2010; Pachur & Olsson, 2012; Persson & Rieskamp, 2009; Wirebring et al., 2018).

However, already B. Brehmer (1994, p. 152) suggested that "judgments result from a compromise between rules and specific memories of earlier outcomes". According to Herzog and von Helversen (2018), the blending of both types of processes makes sense from a mere normative and ecological perspective. By pooling together the different forms of information used by both processes, more accurate judgments can be produced than by any process alone. This idea of a mixture or blending of processes is also common in many models in the categorization literature (so called hybrid models), which differ in the way rule and exemplars are combined to make a categorical judgment (e.g., Anderson & Betz, 2001; Love et al., 2004; Nosofsky et al., 1994). For example, the ATRIUM model (Attention To Rules and Instances in a Unified Model, Erickson & Kruschke, 1998) assumes that rule- and exemplar-processes work parallel and independently. The final categorization response is based on a linear combination of the predictions of each process, where the relative weight of each process varies for each stimulus, based on an error-driven learning process. This assumption resonates well with empirical evidence, showing that the similarity of specific exemplars influences the categorization accuracy or speed of new stimuli, even when simple and perfectly predictable rules are available to the participants (either through extensive training or even through explicit instructions) and even if the reliance on similarity may be even detrimental to categorization performance (Allen & Brooks, 1991; Brooks & Hannah, 2006; Hahn et al., 2010; Hannah & Brooks, 2009; Regehr & Brooks, 1993; Thibaut et al., 2018). These results are often interpreted as an automatic and unintentional activation of exemplars (Hahn et al., 2010, see also Macrae et al., 1998). Recently, von Helversen, Herzog, et al. (2014) found similar results in a multiple-cue judgment task where participants had to judge the qualification of six fictitious job candidates based on four cues (education, motivation, skills, and quality of work experience). Even though the overall judgments of new candidates were described well by a rule-based model, the judgments were also influenced by the similarity of learned exemplars: New candidates were judged as more qualified when they resembled highly-qualified exemplars and as less qualified when they resembled less-qualified exemplars (see also Rosner & von Helversen, 2019). The results also showed that the effect of similarity was larger

when participants reported to rely more on visual appearance (i.e., similarity of job candidates), suggesting that the reliance on exemplars can be deliberate and strategic. These results show that both types of processes are used simultaneously, even when the environment and the task encourage a purely rule-based processing.

Therefore, a more reasonable assumption about the integration of both processes is that people base their judgments simultaneously on rules and exemplars by blending these two processes together, but depending on the actual stimuli, the environment, or task structure, the relative contribution of each process might differ (e.g., Albrecht et al., 2020; Bröder et al., 2017; Wirebring et al., 2018). Based on these considerations, Bröder et al. (2017) proposed the *RulEx-J* model to measure the relative contribution of each process over a series of trials.

The RulEx-J Model

Like ATRIUM (Erickson & Kruschke, 1998), the RulEx-J model assumes that rule- and exemplar-based processes work in parallel and independently. The final judgment is then a blending of the preliminary judgments of both distinct processes. Thus, the RulEx-J model is more in line with the empirical evidence presented before and allows a more fine-grained measurement of the judgment processes involved (Bröder & Gräf, 2018; Bröder et al., 2017). The conceptual idea of the RulEx-J model is depicted in Figure 2.

In the RulEx-J model, stimuli are represented as vectors of features (e.g., Hintzman, 1984; Love et al., 2004; Nosofsky, 1984). For example, the fictitious flower in Figure 2 is represented by four binary features. The to-be-judged stimulus (i.e., the probe p) is processed by two modules, an exemplar module (E) and a rule module (R). The preliminary judgment of the exemplar module J_E is based on the similarity of the probe to the stored exemplars. In contrast, the preliminary judgment of the rule module J_R is based on the integration of cues according to an abstracted rule.

In the articles using the RulEx-J model, the exemplar- and rule-based processes in the corresponding modules are modeled according to the formal models in Equations (1) - (2) (i.e., the context model extend to numerical judgments and the cue abstraction model, Bröder and Gräf, 2018; Bröder et al., 2017, see also Manuscript I of this thesis). However, other implementations of the respective judgment processes could also be used. For instance, in Manuscript III of this thesis, I used the GCM (Nosofsky, 1984, 2011) as a formal model of the exemplar module. The fi-

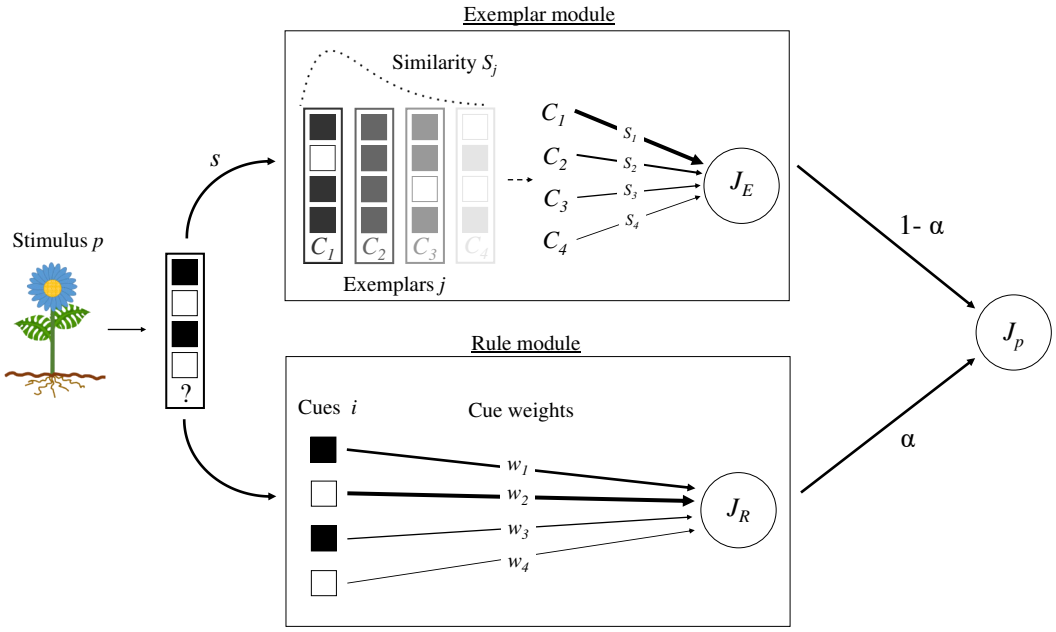
nal judgment J_p is a combination of the results of both processes, weighted by the parameter α (see Figure 2):

$$J_p = \alpha \times J_R + (1 - \alpha) \times J_E \tag{4}$$

The α parameter measures the relative contribution of the rule- and exemplar-based process on the final judgment. It can range from 0 to 1, with larger values indicating more rule-based processing.

Figure 2

Schematic representation of the RulEx-J model



Note. A to-be-judged probe p, consisting of four binary features, is processed by the exemplar module (upper half) and the rule module (lower half) in parallel. In the exemplar module, the probe is compared to each of four exemplars stored in memory and the criterion values C_j of each exemplar j is weighted according to the similarity of the exemplar and the probe (S_j) to produce the judgment J_E. Larger arrows indicate a higher similarity and thus larger influence on the final judgment. In the rule module, the probe is decomposed into separate cues which are then integrated according to a rule to generate the judgment J_R. Larger arrows indicate higher cue weights. Both interim judgments of the rule- and exemplar-module are then weighted by α and (1-α) respectively and then integrated into a final judgment J. Figure is based on Figure 1 in Bröder et al. (2017).

The implementation of the integration of both processes in Equation (4) assumes a constant continuous blending of both processes, which is only one (rather simple) possible implementation of the possible blending or mixture processes (Albrecht et al., 2020; Anderson & Betz, 2001; Erickson & Kruschke, 1998; Schlegelmilch et al.,

2021).

In three experiments, Bröder et al. (2017) positively tested the validity of the α parameter by showing that direct strategy instructions and changes to the learning task had the intended effects on the dominant type of processing and that these changes were adequately captured by the parameter α. In addition, they found that, overall, the RulEx-J model was better in predicting new judgments of participants than each of its sub-modules (i.e., pure exemplar or rule model). However, the RulEx-J model is (so far) mostly intended as measurement model which provides "more sophisticated or sensitive measures of processing" (Bröder et al., 2017, p. 504) and not as an epistemic model (but see Albrecht et al., 2020).

Despite the success and progress of the multiple-cue judgment research program in investigating the processes underlying people's judgments, in the following chapter I highlight and solve several issues and shortcomings in the way the here presented models are currently used. By making the RulEx-J model measure the relative contribution of rule- and exemplar-based processes more accurately and more robust (Manuscript I), by demonstrating and solving problems brought by the uncritical adaptation of exemplar models and the experimental paradigm from categorization research (Manuscript II), and by showing how to use these models of quantitative judgments not only on simple and artificial but also on realistic and complex stimuli (Manuscript III), the manuscripts presented in my thesis improve upon the presented models of quantitative judgments, the way they are currently used, and extend their possibilities. Therefore, this thesis not only adds to the methodological development and application of cognitive models in multiple-cue judgment research but also provides possibilities for future theoretical insights into the underlying judgment processes and for additional areas of application of the formal models of these processes. In the following chapter, I summarize the three manuscripts and present their core results (the full manuscripts can be found in Appendix D).

3 Improving and Extending Models of Quantitative Judgments

In three manuscripts presented in this chapter, I highlight and solve different problems of how models of quantitative judgments are currently implemented and used in multiple-cue judgment research. Specifically, in Manuscript I I implement and test a hierarchical Bayesian version of the RulEx-J model which improves parameter estimation and comparison of latent parameters between conditions. In Manuscript II, I highlight a major problem of how exemplar-based models are used in multiple-cue judgment research and present a new extended exemplar model which resolves this issue. In Manuscript III, I show how the existing models of numerical judgments can be used to model judgments of non-artificial complex stimuli where the cues representing these stimuli and needed for the judgment models are not known beforehand.

3.1 The hierarchical Bayesian RulEx-J Model

Izydorczyk, D., & Bröder, A. (in press). Measuring the mixture of rule-based and exemplar-based processes in judgment: A hierarchical Bayesian approach. *Decision.*

As mentioned in the previous chapter, Bröder et al. (2017) proposed the RulEx-J model as a more sensitive and theoretically more adequate method to measure the relative contribution of rule- and exemplar-based processes in people's judgments than classifying individuals based on the best-fitting model. So far, articles which used the RulEx-J model employed maximum-likelihood or least-squares (LS) optimization to estimate model parameters (Albrecht et al., 2020; Albrecht et al., 2021; Bröder & Gräf, 2018; Bröder et al., 2017). However, Bröder et al. (2017) reported that using these estimation approaches, the α parameter tends to be biased towards one (i.e., rule-based processing) when there is a lot of noise in the data, because the cue-abstraction model (reflecting the rule module) has more free parameters

than the context model for continuous judgments (reflecting the exemplar module) and is thus prone to overfitting. This is a major drawback of the RulEx-J model, since most studies in the multiple-cue judgment literature are interested in the factors that influence the main mode of processing, but the bias towards rule-based processing could lead to wrong conclusions. For instance, a manipulation that only effects the levels of noise in the responses of participants, but has no affect on the underlying cognitive process, would still show a difference in processing as captured by the RulEx-J model.

In this first manuscript, we proposed a hierarchical Bayesian implementation as a potential solution to this problem. The hierarchical Bayesian modeling framework has become very popular in cognitive psychology, since it offers many advantages over maximum-likelihood or least-squares approaches (for introductions, see Lee & Wagenmakers, 2014; McElreath, 2020; Rouder et al., 2018). For instance, the partial pooling of information enabled by the hierarchical structure of the model often leads to more accurate parameter estimates (e.g., Farrell & Ludwig, 2008; Katahira, 2016; Rouder et al., 2005). Although, hierarchical models are not constrained to a Bayesian framework (e.g., Erdfelder, 1993), it is more flexible and allows a more straightforward implementation of even complex hierarchical models (Lee, 2018). Further, for our specific case we expected the hierarchical Bayesian approach useful in two additional ways. First, the Bayesian method should automatically control for the different complexities of the exemplar- and rule-module, since the blending of the RulEx-J model is similar to a model selection between the exemplar and rule modules (Lee, 2008). This should lead to a reduction of the problematic bias towards rule-based processing as reported by Bröder et al. (2017). Second, the advantages of a hierarchical Bayesian approach for estimating and testing hypothesis about latent parameters are especially pronounced when the number of trials per participant is small (Böhm et al., 2018; Katahira, 2016; McElreath, 2020), which is a common situation in most multiple-cue judgment studies where there are often only 16 data points available per participant (e.g., Bröder & Gräf, 2018; Juslin et al., 2003; Pachur & Olsson, 2012).

In the first part of the manuscript, we thus developed and tested the hierarchical Bayesian RulEx-J model. Like the original RulEx-J model, we used the cue-abstraction model (Juslin et al., 2003) and the context model (Juslin et al., 2003; Medin & Schaffer, 1978) presented in the previous chapter in Equations (1) - (3) as respective models of the rule- and exemplar modules of the RulEx-J model. We

also assumed that the response y of a participant i in trial t is normally distributed around the weighted average of the rule-based and an exemplar-based process according to Equation (4), with some person specific precision σ (i.e., $y_{it} \sim N(J_{it}, \sigma_i)$).

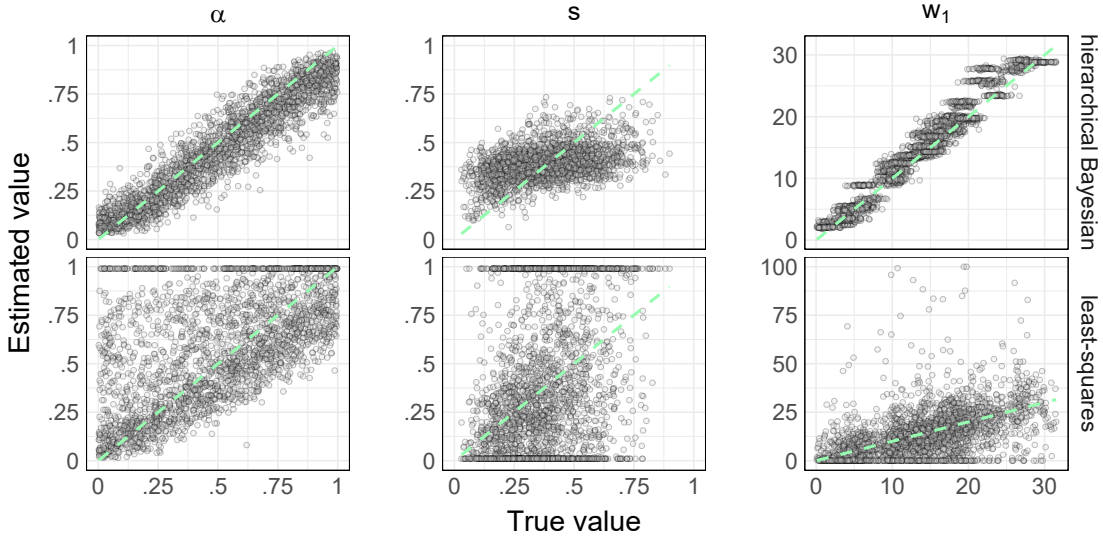
In computer simulations, we tested the ability of the hierarchical Bayesian RulEx-J model to recover parameters and its robustness against different magnitudes of noise in the data. For this purpose, we simulated judgment data of multiple synthetic participants and added normally distributed error with mean $\mu = 0$ and different standard deviations of $\sigma_\epsilon = 0$, 2, 4, 8 to the generated judgments, in order to simulate different levels of noise (none, low, medium, high). The simulations allowed us to check whether the model is correctly implemented, but also how the model behaves under more realistic conditions, and whether it was indeed more robust against noise in the data than the LS-approach.

The results showed that the hierarchical Bayesian implementation of the RulEx-J model is able to accurately recover the underlying parameter values when the simulated judgment data was noise free (i.e., $\sigma_\epsilon = 0$), indicating a correct implementation of the model. Further, as expected the results showed that, compared to the original LS-approach, the hierarchical Bayesian approach led to more accurate and less biased estimates when there were high levels of noise in the data (i.e., $\sigma_\epsilon = 8$, see Figure 3). The results also demonstrated that the misestimation behaviour of the hierarchical Bayesian implementation under high levels of noise is more reasonable than the behaviour of the LS-approach. As evident in Figure 3, besides being less accurate in general, the LS-estimates tended towards the upper or lower parameter boundaries independent of the true value. For instance, for the α parameter we replicated the bias towards rule-based processing mentioned by Bröder et al. (2017), where the parameter was estimated to be one in some instances. The hierarchical Bayesian estimates on the other hand show clear signs of shrinkage, were the estimates are pulled towards the sample mean. This shrinkage is especially strong for the s and the cue weight parameters. Although these estimated individual parameters are still not recovered with high accuracy, the way in which the estimation fails is thus more reasonable than the erratic behaviour of the LS-estimates and inference about group-levels parameters are still possible. In addition, the simulation results also highlighted certain limitations for multiple-cue judgment researchers on what inferences they might be able to draw from their data. Given the realistic number of participants, number of trials per participant, stimuli structure, and levels of noise, the results indicate that the available data are often not informative enough to esti-

mate individual level parameters accurately given the observed amount of shrinkage. Therefore, researchers should focus on making inferences on the group-level.

Figure 3

Parameter recovery for high levels of noise (i.e., $\sigma_\epsilon = 8$, Izydorczyk and Bröder, in press)



Note. Each dot represents a single synthetic participant.

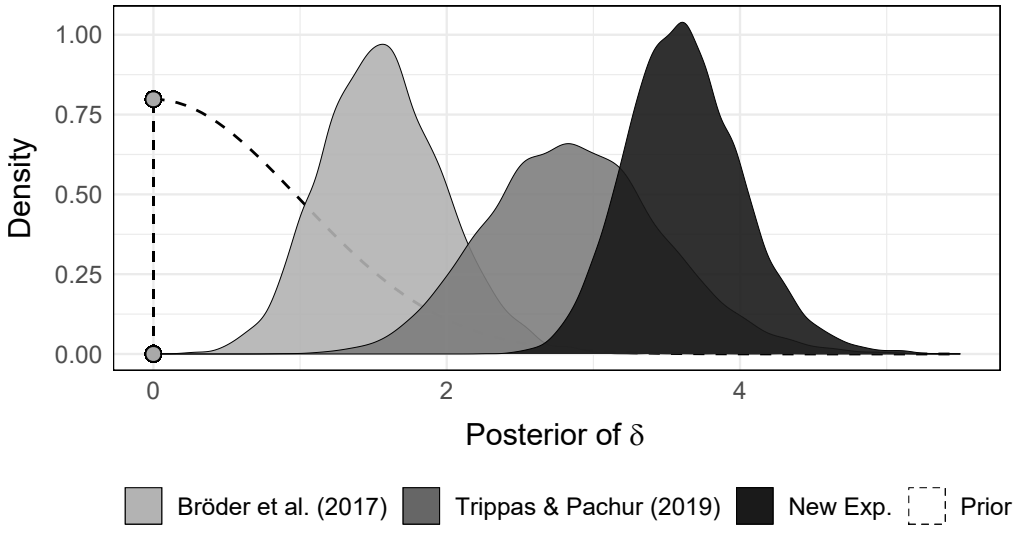
In the second part of the manuscript, we applied the hierarchical Bayesian model to three data sets: Experiment 1B from Bröder et al. (2017, $N = 60$), Experiment 1 from Trippas and Pachur (2019, $N = 60$), and one new preregistered experiment ($N = 238$). Each experiment used different stimuli, judgment criteria, and different manipulations to influence the dominant type of processing. The new experiment differed from the traditional multiple-cue judgment experiments in so far as it did not include a learning phase, in which participants usually learn the exemplars, their criterion values, and the relationship between cues and criterion values. Instead, participants directly received different information to make their judgments, depending on the condition. More specifically, participants had to judge the price of 16 fictitious flowers on a scale from 0 to 100 €. The flowers could differ on four binary cues: petal color (red/blue), leave form (thin/thick), petal form (round/star-shaped), and root form (shallow/thick). In the exemplar condition, participants saw the same selection of eight flowers and their criterion values. Participants were informed that all 16 flowers could be judged based on their similarity to these exemplars. In the rule condition, participants were informed about the approximate

price of the cheapest flower. Further, we provided ranges of possible price increases for each feature. For instance, participants were told that star-shaped flowers cost 10 to 20 € more than round flowers. All participants judged all 16 flowers twice in randomized order, resulting in 32 trials per participant. We expected that in each data set, the α parameter should be higher in the respective rule condition compared to the exemplar condition, indicating more rule-based processing.

Following the recommendation of Böhm et al. (2018), who showed that ignoring the hierarchical structure in the data or using individual-level estimates from a hierarchical model in a subsequent test leads to wrong conclusions, we directly implemented the difference in α between conditions in the hierarchical Bayesian RulEx-J model. The resulting posterior distribution of the δ parameter, which reflects the standardized mean difference in α between both groups, showed that as expected the α parameters were on average larger in the rule conditions than in the exemplar conditions in all (re)analysed experiments (i.e. $\delta > 0$), indicating more rule-based processing (see Figure 4).

Figure 4

Prior and posterior distributions of δ for three different experiments (Izydorczyk & Bröder, in press)



Note. The effect size parameter δ reflects the mean difference in α between the rule/lbc and exemplar/dcl condition in each of the three experiments on a standardized scale, where $\delta > 0$ indicates higher α values and thus more rule-based processing in the rule condition.

Overall, in this manuscript we developed, tested, and applied a hierarchical

Bayesian implementation of the RulEx-J model, which provided more accurate and robust parameter estimates than the original LS-approach. Furthermore, by directly implementing differences in α between conditions into the model, we were able to reproduce results of two previous experiments and provided additional evidence for the validity of the mixture parameter α with a new experiment. The well document and openly available scripts and model files thus provide the research community with a tested tool to measure and compare the blending between rule- and exemplar-based processes.

3.2 Problematic Procedure for Estimating Parameters in Exemplar Models of Quantitative Judgments

Izydorczyk, D., & Bröder, A. (2021). Exemplar-based judgment or direct recall: On a problematic procedure for estimating parameters in exemplar models of quantitative judgment. *Psychonomic Bulletin & Review*, *28*, 1495–1513.

In the second manuscript, we highlighted a major problem in the way exemplar-based models are used in almost all multiple-cue judgment studies, beginning with their introduction to this area of research by Juslin and colleagues (2003, 2002).

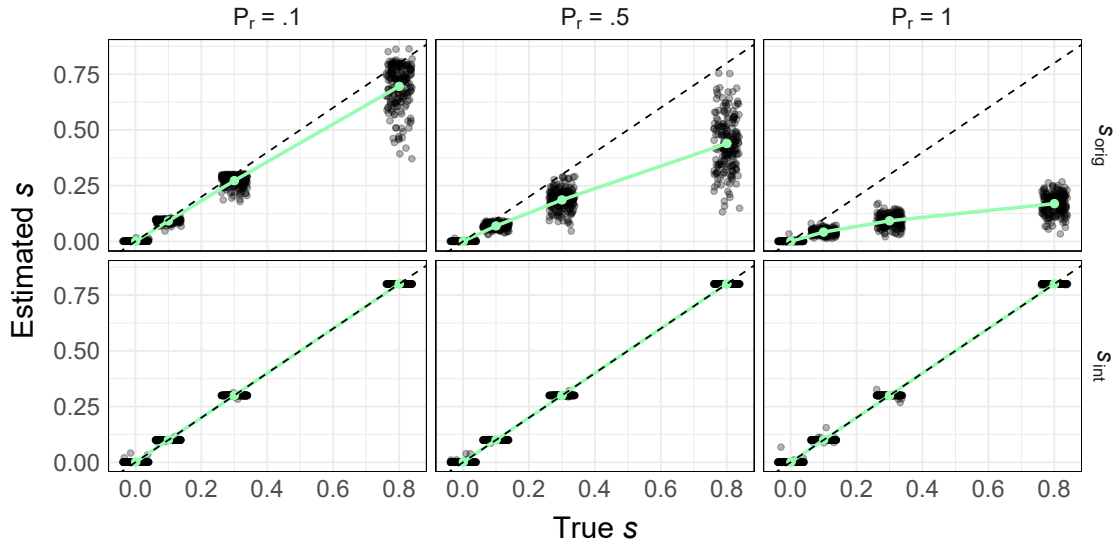
The typical experiment in the multiple-cue judgment literature involves an extensive training phase where participants are repeatedly presented with the same set of few well distinguishable exemplars, so they can learn their cues, criterion values, or the rule connecting the cues to the criterion values (e.g., Bröder & Gräf, 2018; Hoffmann et al., 2013; von Helversen & Rieskamp, 2009a). For instance, in Bröder and Gräf (2018) participants learned to judge eight possible exemplars over a series of eight blocks. After the training phase, participants were able to judge on average 78% of the exemplars correctly and 46.67% of the participants had learned all exemplars perfectly. This experiment is by far no exception, since many studies either implement a learning criterion that participants have to reach to finish the training phase (e.g., 85% correct in Trippas & Pachur, 2019) or include an even longer training phase (e.g., up to 40 blocks in Wirebring et al., 2018), to make sure that participants are able to learn all exemplars. Model parameters are then estimated either on the last blocks of the training phase or on the data of the judgment phase.

In this second manuscript we proposed that the judgments of participants in these later stages of the experiment are a mixture of two qualitatively distinct cognitive processes: Judgment or direct recall. When participants have to judge a trained exemplar there are two possible ways participants can respond. They either have learned the exemplar and its criterion value and thus can directly retrieve the criterion value from memory, or they have not learned the exemplar and thus have to judge its criterion value, as if it was a new stimulus. Using the context model in Equations (2) - (3) as an example (with one general s parameter), we predicted

that neglecting this distinction leads to impaired estimation and validity of model parameters, resulting in a decreased model fit in general. Specifically, we predicted that the s parameter will be biased towards 0 if all data points (correctly recalled exemplars and other trials) are jointly used to estimate the model parameters. The rationale behind this prediction can be demonstrated with a short example.

Figure 5

Parameter recovery results (Izydorczyk & Bröder, 2021)



Note. The figure shows the parameter recovery results of the s parameter of the context model, for different true values of s and different probabilities of recalling exemplars correctly (P_r). Each black dot represents one synthetic participants. Green dots represent the corresponding mean for a specific true value of s and P_r. s_{orig} and s_{int} are the estimated s parameters using the traditional approach (i.e., not differentiating between recalled trained exemplars and other stimuli) or the latent-mixture exemplar-model, respectively.

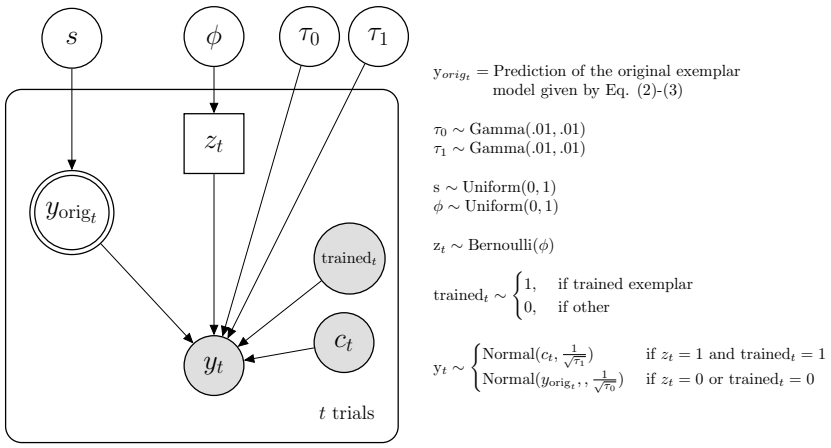
Suppose there are only two exemplars which vary on only one binary cue, $a = [0]$ and $b = [1]$. Their criterion values are 42 and 7, respectively. Like a, the probe is $p = [0]$ and according to Equation (2) the similarity of the probe to the two exemplars is thus $S_a = 1$ and $S_b = s$. The judgment of our fictitious participant of the probe is 42. When $s = 1$, the exemplar model predicts the mean of the criterion values of all exemplars, which in this case is 24.5. However, when $s = 0$ the model predicts 42, which is the criterion value of a. Thus, if there is an identical exemplar to the probe and the judgment of the probe is equal to the criterion value of this identical exemplar, the best model fit is achieved when $s = 0$. Thus, when the responses of participants include correctly recalled exemplars (i.e., the judgment of a trained

exemplar is identical to the actual criterion value of this exemplar), the estimated s parameter will be biased towards 0 and the bias will be larger the more correctly recalled exemplars there are in the data.

The results of our simulated judgment data based on the exemplar model with different values of s (.001,.1,.3,.8) and different probabilities of recalling an exemplar correctly (P_r), confirmed our prediction (see top row of Figure 5). When the s parameter was estimated based on all data points (i.e., the original approach, s_{orig}), the s parameter was increasingly biased towards 0, the more correctly recalled exemplars there were in the data.

Figure 6

Graphical model representation of the latent-mixture exemplar model with integrated direct recall (Izydorczyk & Bröder, 2021)



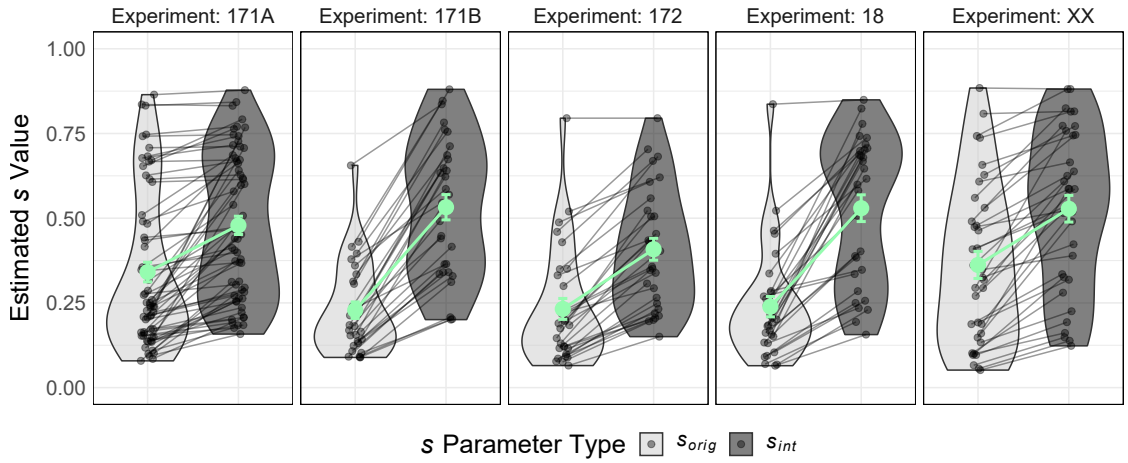
Note. Observed variables are depicted as shaded nodes, unobserved variables as unshaded nodes, discrete variables as square nodes, continuous variables as circular nodes, deterministic variables as double-bordered nodes, and stochastic variables as single-bordered nodes. c_t is the criterion value of the exemplar in trial t.

As one potential solution to this problem, we proposed an extended version of the exemplar model, which incorporates the possibility of recalling exemplars directly. The model depicted in Figure 6 assumes that when the probe in trial t is a trained exemplar, participants can either directly recall the criterion value c of this exemplar with a probability ϕ, or they do not recall the exemplar and judge it based on the original exemplar model. As intended, the estimated s parameter of this latent mixture extended exemplar model with integrated direct recall (s_{int}) was unbiased and independent of the proportion of directly recalled exemplars (see bottom row of Figure 5).

Following up on the simulations, we reanalysed data from five different experiments in order to test whether the effects found in the simulation also extend to empirical data. In all data sets, participants had a high proportion of correctly recalled exemplars at the end of the training phase (average P_r ranging from .43 to .85). Therefore, we predicted that the s parameter estimated with the original exemplar model (s_{orig}) will be lower than the unbiased s parameter estimated with the latent-mixture extended exemplar model (s_{int}). The results shown in Figure 7 confirmed this prediction. In each of the five data sets, the s_{orig} parameter was lower than the s_{int} parameter. Additional analysis showed that the new extended exemplar model also provided a better fit to the data in all reanalysed experiments.

Figure 7

Results of the reanalysis of five different experiments (Izydorczyk & Bröder, 2021)



Note. The figure shows the median posterior values of s_{int} and s_{orig} for each participant and for each data set. s_{orig} and s_{int} are the estimated s parameters using the traditional approach (i.e., not differentiating between recalled trained exemplars and other stimuli) or the latent-mixture exemplar-model, respectively. Green dots represent the means and the corresponding standard errors.

In summary, the results of Manuscript II highlighted problems with the application of exemplar models in multiple-cue judgment research. We showed that neglecting to differentiate between learned exemplars and judged stimuli led to biased parameter estimates, impaired validity of model parameters, and a decrease of model fit. This finding is problematic since in many multiple-cue judgment studies participants are classified as exemplar or rule users, depending on which model fits better to the data. Thus, in some instances the artificially decreased fit of the exemplar model could lead to wrong conclusions about the process which participants

used to make their judgments. However, it should be noted that the highlighted issues are not due to a problem with the exemplar model itself, but the combination of the model and the adaptation of the experimental paradigm from categorization research which involves having few well-distinguishable stimuli. The main difference between the categorization and judgment experiments, and the reason why this problem is not (or less) noticeable in categorization research, is the scale of the criterion value. Whereas the criterion in categorization studies is categorical and multiple stimuli share the same criterion value (e.g., in Shepard et al., 1961, four stimuli belong to Category 1 and four in Category 2), the criterion value in judgment studies is continuous and most exemplars have a unique criterion value (e.g., in Trippas and Pachur, 2019, only one training exemplar has the value .80) and this unique exemplar-criterion mapping leads to the strong bias in s.

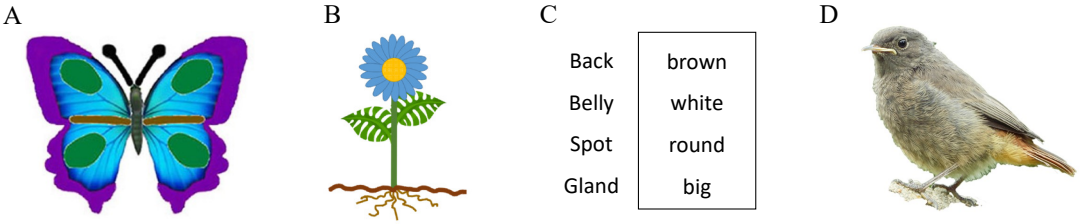
3.3 Modeling Numerical Judgments of Realistic Stimuli

Izydorczyk, D., & Bröder, A. (2022). *What is the airspeed velocity of an unladen swallow? Modeling numerical judgments of realistic stimuli.* [Manuscript submitted for publication]. Department of Psychology, University of Mannheim.

In their overview of the history of JDM research, Goldstein and Hogarth (1997) proposed that one of the major questions relevant to the future of JDM research is to what extent one can generalize from laboratory studies with abstract tasks and artificial stimuli to behavior in the real world. In contrast to the traditional judgment research around the SJT, studies investigating the underlying cognitive processes of multiple-cue judgments so far exclusively relied on rather simple artificial stimuli such as depicted in Figure 8A-C. One of the main reasons for that is that the cues and cue values of the respective judgment objects need to be known in order to use the cognitive models in this line of research (and presented in this thesis). For instance, whereas the similarity in exemplar models is based on the similarity of the cue values of the exemplars and the probe, in rule-based models the cues are directly integrated according to some rule. However, for complex objects encountered in the real world the cues people use to make their judgments are often unknown and thus it is not possible to use the models on judgments of these complex objects.

Figure 8

Example stimuli used in different multiple-cue judgment experiments

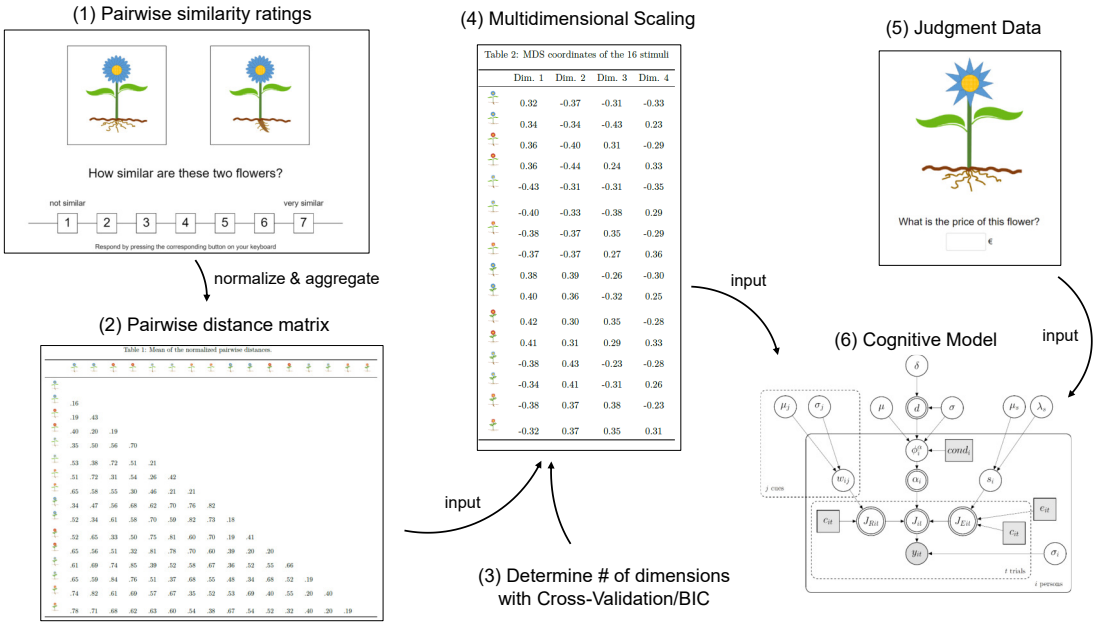


Note. The stimuli are: A) Hoffmann et al. (2018), consisting of 4 cues; B) Izydorczyk and Bröder (in press), consisting of 4 cues; C) Trippas and Pachur (2019), consisting of 4 cues; D) Izydorczyk and Bröder (2022).

In this third manuscript we showed how cues and cue values can be generated for complex natural stimuli, where the cues are not known beforehand, and that these generated cues can then be used in computational models of numerical judgments. The general approach is depicted in Figure 9. Based on early categorization research (e.g., Nosofsky, 1992; Shin & Nosofsky, 1992), we proposed that cues and cue values can be extracted from pairwise similarity ratings using multidimensional scaling analysis (MDS, Kruskal, 1964; Shepard, 1962, Steps 1-4 in Figure 9). In MDS, objects are represented as points in a multidimensional space, where similar objects are located closer together than dissimilar objects (Hout et al., 2013; Shepard, 1962). The dimensions of the MDS solution can then be used as cues in cognitive models to analyse the responses of participants (Steps 5-6 in Figure 9).

Figure 9

General procedure used in Izydorczyk and Bröder (2022)



Note. The figure shows the general procedure used in both studies presented in Izydorczyk and Bröder (2022). pairwise similarity ratings (Step 1) are aggregated and transformed into a pairwise distance matrix (Step 2), which is then used in a subsequent multidimensional scaling analysis (Step 3 & 4). The resulting MDS dimensions can then be used as cues in cognitive models modeling data from judgment experiments using the same stimuli (Step 5 & 6).

By using artificial stimuli with a known cue-structure and existing data of a previous judgment experiment, the first study reported in this manuscript served as a proof-of-concept for the general approach depicted in Figure 9. For this purpose, we

used the 16 artificial flower stimuli and the data from the corresponding experiment presented in the first manuscript (Izydorczyk & Bröder, in press). The flowers were generated by combining four different binary features: Petal color (red/blue), leave form (thin/thick), petal form (round/star-shaped), root form (shallow/thick), see Figure 8B for an example. In the judgment experiment, participants were provided with different information in the two conditions, which fostered either exemplar- or rule-based processing. In this first study, we expected to recover the underlying cue structure of the artificial stimuli (i.e., four cues). Further, we expected that the analysis of the judgment experiment using the hierarchical Bayesian RulEx-J model should yield similar results when the MDS-generated cues are used instead of the experimentally defined cues as in Manuscript I. That is, the α parameter should be higher in the rule condition than in the exemplar condition, indicating more rule-based processing.

The results confirmed both predictions. Based on pairwise similarity ratings from $N = 40$ participants, the attributes generated by the MDS analysis fully recovered the underlying cue structure. In addition, using the MDS-cues in the hierarchical Bayesian RulEx-J model lead to very similar posterior predictions for the individual trials ($r(7614) = .99$, $p < .001$) and the same inference regarding the difference in the α parameter between conditions, that is, higher average α values in the rule condition than in the exemplar condition.

In the second study, we replicated an experiment of Pachur and Olsson (2012) and Trippas and Pachur (2019) showing that the type of learning task and feedback during the learning phase impacts whether participants relied more on exemplar-based processing or rule-based processing. However, instead of the usual artificial stimuli, we used natural complex stimuli (i.e., pictures of birds, see Figure 8D for an example) with an a priori unknown cue structure. We again first collected pairwise similarity ratings for the $K = 32$ stimuli from $N = 97$ participants (Step 1 in Figure 9). The subsequent MDS analysis (Steps 2-4 in Figure 9) showed that, according to a cross-validation procedure, the similarity ratings of the bird images were best described by three dimensions.

In the judgment experiment, $N = 78$ participants had to judge the flight speed of the 32 different birds, of which $k = 12$ were presented as exemplars during the training phase. As in the original experiments of Pachur and Olsson (2012) and Trippas and Pachur (2019), participants either learned by comparison or by direct criterion learning. In the learning by comparison condition, participants were asked

to decide which of two birds is the faster bird and received feedback about the correct answer. In the direct criterion learning condition, participants had to judge whether a bird can be classified as a slow or fast bird. Beside the correctness of their answer, they received feedback about the actual flight speed of the presented bird. As in the first study, we analyzed the data with the hierarchical Bayesian RulEx-J model using the MDS-generated attributes as cues. Results showed that we successfully replicated the results of Pachur and Olsson (2012) and Trippas and Pachur (2019), in that participants showed more rule-based processing in the learning by comparison condition than in the direct criterion learning condition.

Overall, this manuscript demonstrated how cognitive models of numerical judgments can be used on natural complex stimuli with an unknown cue structure. This extends the possibilities of these judgments models in two important ways. First, as stated by Goldstein and Hogarth (1997), we are able to test whether the many laboratory findings about the cognitive processes of multiple-cue judgments generalize to more complex stimuli as encountered in the real world. Second, this allows us to use the developed and well tested models of people's judgment process, and the knowledge from the corresponding research, to applied problems in people's everyday life, which in general involve complex judgment objects (e.g., estimating the carbon footprint or nutritional values of food items).

4 General Discussion

In my dissertation, I improved, developed, and extended models of quantitative judgments. Specifically, I implemented and tested a hierarchical Bayesian version of the RulEx-J model, developed a latent-mixture extended exemplar model better suited to the experimental paradigm, and extended the scope of the judgment models presented throughout this thesis by demonstrating how they can be used on complex stimuli with initially unknown cue structure.

In Manuscript I, we developed, tested, and applied a hierarchical Bayesian version of the RulEx-J model. The simulation results showed that the hierarchical Bayesian approach led to more accurate and robust parameter estimates than the hitherto used method. Further, the results highlighted the (non-) informativeness of the typical multiple-cue judgment experiment in which the data are often not informative enough to accurately estimate individual-level parameters. Finally, using the hierarchical Bayesian RulEx-J model we were able to reproduce results of previous experiments and provided evidence for the validity of the α parameter in a new experiment. In summary, this manuscript provided researchers with an improved state-of-the-art method to estimate and test hypotheses about the relative contribution of rule- and exemplar-based processes in people's judgments.

In Manuscript II, we demonstrated through simulations and reanalysis of five experiments that the combination of exemplar models and the experimental paradigm used in almost all multiple-cue judgment studies leads to a biased estimation of model parameters and reduced model fit when the difference between directly recalled exemplars and other stimuli is not taken into account. The manuscript also presents a solution to this problem by introducing an extended exemplar model which integrates a direct recall process of trained exemplars. This manuscript thus uncovers and solves a major shortcoming in the way exemplar models are currently used in multiple-cue judgment research.

In the last manuscript, we demonstrated how to model people's judgments of complex and realistic stimuli by extracting the necessary cues from pairwise similarity ratings using multidimensional scaling analysis. Primarily, the results of the four experiments in two studies showed that using this approach, known cue structures and results from previous experiments can be replicated and even generalized to

new complex stimuli. However, the implications of Manuscript III are even broader. In their description of the science of psychology, the American Psychological Association (APA) states that "psychologists apply the understanding gleaned through research to create evidence-based strategies that solve problems and improve lives." (APA, 2013, para. 2). The results presented in Manuscript III potentially enable us to do exactly that: Making it possible to use the knowledge about people's judgment processes, which was gained through experimental research over the last two decades, to investigate and improve real-life judgment problems in a theory- and evidence-based manner. For instance, helping people to estimate the calorie content of food items more accurately might have large potential benefits on people's health (e.g., König et al., 2019; Sacks et al., 2011).

Taken together, the results of the three manuscripts described here contribute to the model-based study of cognitive processes underlying people's judgments. By implementing state-of-the-art methods, improving upon current practices, and broadening the scope of the existing research, the results reported in this thesis add to developing and testing of theories of quantitative judgments.

4.1 Open Questions and Future Directions

Although, the three manuscripts presented in this thesis solved several specific problems or shortcomings of the formal models used in multiple-cue judgment research, they also offer new possibilities or raise new questions for future research. In the following, I discuss some of the remaining open questions and possible future directions.

4.1.1 Investigating the Blending Process

The hierarchical Bayesian RulEx-J model presented in Manuscript I allows to measure the relative contribution of rule- and exemplar-based processes more accurately. While this is an important improvement of the RulEx-J model as a measurement model, it does not improve the epistemological value of the RulEx-J model by further testing or validating the assumed blending process. Many different possible blending processes have been proposed in the categorization and multiple-cue judgment literature, as well as related research fields. For instance, whereas the RulEx-J model assumes a constant continuous blending of both processes over a

series of trials (Bröder et al., 2017), the ATRIUM model (Erickson & Kruschke, 1998) assumes that the relative contribution of each process changes between trials as participants learn which process is best suited for each exemplar. Instead of a continuous blending, the CX-COM[2] model assumes a two-step process (Albrecht et al., 2020). First, only one exemplar is retrieved from memory, where exemplars which are more similar to the probe have a higher chance of being retrieved. The criterion value of this one exemplar is then adjusted based on a rule-based process to produce the final judgment. In contrast, the Rule-Plus-Exception model for category learning (RULEX, Nosofsky et al., 1994) assumes that people use simple rules but revert to specific memorized exceptions to these rules if necessary. Relatedly, the Category Abstraction Learning (CAL) framework assumes that rule-like representations, which emerge from complementary similarity and dissimilarity mechanism, are mainly used for categorization, except for some instances were specific memorized exemplars are used (Schlegelmilch et al., 2021). Thus, one task for future research is to thoroughly test and compare theoretically possible blending processes.

However, although the Bayesian framework used in Manuscript I offers many possibilities to test and compare even complex models (e.g., Radev et al., 2020), this task cannot be solved through modeling alone. Rather, it requires carefully designed experiments and stimulus materials to be able to differentiate between different blending processes, since in the typical multiple-cue judgment experiment only few data points are informative enough to even differentiate between predictions of rule- and exemplar-based models, and probably even less so, between different mixture processes. A potential way of dealing with this problem could be to look at differences in predictions between rule- and exemplar-based processes in other dependent variables and data modalities than only participant's judgments (Glöckner, 2009). One of these possible additional variables are response times, which are often used together with other behavioral variables as outcomes of cognitive models (e.g., Busemeyer & Townsend, 1993; Gaissmaier et al., 2011; Ratcliff & Rouder, 1998). For example, Klauer and Kellen (2018) extended traditional multinomial processing tree models to incorporate response times in addition to response frequencies, which lead to new insights into the ordering of memory-retrieval and guessing processes. In addition, the exemplar-based random walk model (an extension of the GCM, Nosofsky & Palmeri, 1997) predicts not only choice-probabilities but also response

[2]combining Cue abstraction with eXemplar memory assuming COMpetitive memory retrieval

times and has been shown to account for a variety of classification response time patterns, such as familiarity or practice effects (Nosofsky & Palmeri, 1997). However, although there are theories which predict response times for rule-based classification models of simple logical choice rules (e.g., Nosofsky & Little, 2010), so far there are no good theories about response-time predictions for the rule-based models used in judgment research. Based on research on decision-making, one prediction would be that the response time should depend on the number of cues (Glöckner, 2009; Payne et al., 1988). Further, the complexity of the rule should also influence response time (assuming that participants use something other than a compensatory linear additive rule, e.g., Bröder and Gaissmaier, 2007).

Besides investigating the type of the mixture process, another open question is how people actually learn the relative importance of rule- and exemplar-based representations in quantitative judgments or how the relative contribution of these processes changes over time. Based on the empirical findings in Chapter 2, one potential approach could be to assume that the weighting or selection of processes depends on the ability to abstract rules (and cue weights) or to form strong representations of specific exemplars. Other approaches based on the theories of category could involve error-driven learning as in the ATRIUM model (Erickson & Kruschke, 1998) or self-confirmatory attention learning (e.g., Schlegelmilch et al., 2021).

4.1.2 Direct Recall and Now What?

Manuscript II highlighted a problem of exemplar models in multiple-cue judgment research, where the combination of experimental paradigm and model led to a systematically biased estimation of parameters. However, the manuscript focused exclusively on exemplar models and thus it is unclear how directly recalled exemplars affect the parameters of other models common in the multiple-cue judgment literature, such as rule-based models or blending models like the RulEx-J model.

In most studies, the criterion values of the exemplars are deterministically defined by the rule in the environment with no (e.g., Hoffmann et al., 2016; Trippas & Pachur, 2019) or only few (e.g., Bröder & Gräf, 2018) exceptions to this rule. Recalling the criterion values of trained exemplars would presumably lead to the estimated cue weights being closer to the cue weights of the environment, since the recalled criterion values perfectly reflect the rule. Thus, even if a participant would guess the criterion values of all new transfer stimuli, but has memorized all training

exemplars and their corresponding criterion values, the estimated cue weights of a rule-based model could still accurately reflect the cue weights of the environmental rule, leading to an overestimated fit of the rule-based model. In addition, given that the fit of rule-based model is (presumably) increased and the fit of exemplar-based models is decreased (as shown in Manuscript II), the mixture parameter of the RulEx-J should be biased towards more rule-based processing when there are many correctly recalled exemplars. Therefore, one question for future research is to investigate the effects of directly recalling exemplars on the parameters of other models common in the multiple-cue judgment literature, as well as an extensive reanalysis of available data sets of multiple-cue experiments to determine the implications for previous results in this line of research, since the prominence of rule-based processing might be an artefact of the experimental paradigm.

On a more theoretical level, Manuscript II raises the question of the role of memory and recall processes in models of multiple-cue judgment tasks in general. In the previous paragraph and in Manuscript II in general, the direct recall of exemplars is treated and viewed as a contamination process akin to random guessing, that is, a psychological processes different from the one intended as the object of interest (Zeigenfuse & Lee, 2010). However, it has been shown that exemplar models (the context model and the GCM) are able to accurately account for data from identification, old/new recognition, or cued recall tasks (Estes, 1994; Nosofsky, 1986, 1988, 1991, 1992; Nosofsky et al., 1989), which corresponds to what we coined direct recall in Manuscript II (i.e., recalling a criterion value when cued with an exemplar). The ability of exemplar models to also account for other types of tasks and responses besides mere categorization or judgment is generally a strength of these models (Estes, 1994; Nosofsky, 1992). Thus, the ability of the context model to account for the direct recall of learned exemplars by setting the $s = 0$, is not a bug, but a feature, even though it leads to an overall biased estimate. Further, in contrast to exemplar models, a pure rule-based processing as outlined in Chapter 2 entails the assumption that individual exemplars are only temporarily stored in working memory in order to abstract the cue weights, but not in long-term memory where only the rule and the corresponding cue weights are stored. Thus, a participant who exclusively relies on a rule-based process should not be able to recall individual exemplars later in the experiment, since they are not encoded and thus cannot be retrieved. The finding that participants who perfectly follow a rule-based strategy still correctly respond to learned exceptions to the rule, might

already be interpreted as evidence for a proposed mixture or blending between process as outlined in Chapter 2, where learning and responding to the exception exemplar is driven by an exemplar-based process. Therefore, a second open question is whether the cued recall of trained exemplars is indeed a contamination process or rather part of the (exemplar-based) judgment processes itself and thus, whether the conceptualization of the "true" s parameter independently of the number of correctly recalled exemplars is appropriate.

4.1.3 Cognitive Models of Numerical Estimation and Seeding Effects in Real-World Contexts

Manuscript III demonstrated how computational models of numerical judgments can be used on complex stimuli with an unknown cue structure. However, although we claim that the demonstrated approach allows us to use the knowledge from the experimental multiple-cue judgment literature to investigate and improve real-life judgment problems, the examples presented in the manuscript itself are either reanalyses or replications of existing multiple-cue judgment experiments. One actual potential use-case of the approach presented Manuscript III outside of the typical multiple-cue judgment research is the investigation of the mechanisms underlying *seeding effects*.

Together with their framework of quantitative estimation, Brown and Siegler (1993) developed the seeding paradigm as an intervention to improve the accuracy of numerical judgments. Brown and Siegler (1993) proposed that both *metric* and *mapping* knowledge about a domain are necessary for accurate estimates. Metric knowledge refers to distributional information of a criterion, for instance, the mean, range, and skewness of the distribution of country populations, whereas mapping knowledge refers to information about the rank order of objects on the criterion dimension. Results of their initial demonstration of the seeding paradigm showed that the judgment accuracy of country population sizes increased by several magnitudes after a short learning phase, where participants learned the criterion value of some items (i.e., the seeding items). This increase in accuracy was found for the seeding items as well as for transfer items and was mainly due to increased metric knowledge. Additional experiments showed that this improvement in judgment accuracy is still measurable after 4 months (Brown & Siegler, 1996) and can not be explained by mere anchoring (Brown & Siegler, 2001). So far, the seeding proce-

dure has only been applied in few studies on the estimation of national populations (Brown & Siegler, 1993, 1996), between-city distances (Brown & Siegler, 1996), college tuition rates (Lawson & Bhagat, 2002), and nutritional information of food items (Wohldmann, 2013, 2015; Wohldmann & Healy, 2020).

There are many similarities and connections between the seeding and the multiple-cue judgment research presented throughout this thesis. First, the experimental paradigms used to study multiple-cue judgments and seeding effects are very similar. In both, participants are repeatedly presented with a small number of seeds/exemplars, receive feedback about their true criterion values during a training phase, and then have to judge new transfer items in a later test phase. Further, Juslin et al. (2008) stated that the cue abstraction process requires knowledge about the ordering of exemplars (i.e., mapping knowledge), as well as knowledge about the range of criterion values (metric knowledge). In addition, explicitly inspired by the metrics and mapping framework developed by Brown and Siegler (1993), von Helversen and Rieskamp (2008, 2009a, 2009b) proposed the mapping model as an alternative rule-based model to the cue abstraction model proposed by Juslin et al. (2003).

One major difference between the two research branches is their theoretical underpinning. Whereas, the multiple-cue judgment research uses computational models of theorized cognitive process, the metrics and mapping framework is a rather vague verbal theory, which does not allow to make precise quantitative predictions. So far, however, despite the overlap between the two research branches, there were no attempts to actually model seeding effects and the underlying processes with the existing computational models, which would provide a stronger theoretical basis and allow to generate new precise and testable hypothesis. A possible reason for this is one of the few differences in the experimental paradigm: The judgment objects. As stated before, multiple-cue judgment research relies on simple artificial stimuli where the researcher defines the cues, the criterion values, and the rule relating the cues and criterion values. In contrast, seeding studies use real-world stimuli with inherent criterion values and unknown cue structure, which makes it difficult to use the models common in multiple-cue judgment research.

Fortunately, Manuscript III clearly demonstrated how computational models of numerical judgments can be used on stimuli with an unknown cue structure and thus how cognitive models can be used to investigate the underlying mechanism of the seeding effect. Besides building a stronger theoretical foundation, this would

allow us, for example, to generate precise and testable predictions about which exemplars/seeds lead to the highest increase in judgment accuracy. This knowledge can then be used in actual real-world judgment tasks, important for people's life. For instance, people with Type 1 diabetes have to estimate the amount of carbohydrates in their meal to adjust the needed insulin dose. Although an accurate estimation of carbohydrates and dosing of insulin is essential for the health of people with Type 1 diabetes (American Diabetes Association, 2015; Laurenzi et al., 2011), many patients have a hard time doing so (e.g., Bishop et al., 2009; Kawamura et al., 2015). Thus, an efficient, easy-to-implement, and theory-based training method like seeding might be helpful for designing intervention or training programs for diabetics and understanding the cognitive processes behind them could inform efficient procedures.

4.2 Conclusion

In light of the call to build stronger theories by using computational models (e.g., Oberauer & Lewandowsky, 2019) the famous quote "all models are wrong, but some are useful" (G. E. P. Box) may be extended by: "[and] useful models produce better science" (Smaldino, 2019, p. 9). However, models only become and stay useful by constantly adapting them to new theoretical insights, methodological and statistical developments, and empirical findings. The contribution of this thesis was to improve and extend models of quantitative judgments and the way they are currently used. By implementing state-of-the-art methods, improving upon current practices, and broadening the scope of the existing research, the three manuscripts presented in this thesis not only advance the current state of the literature on quantitative judgments, but also contribute to establishing, testing, and applying theories about the processes underlying people's judgments.

5 Bibliography

Albrecht, R., Hoffmann, J., Pleskac, T., 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. https://doi.org/10.1037/xlm0000772

Albrecht, R., Rosner, A., Rieskamp, J., & von Helversen, B. (2021). *Towards understanding the within-trial dynamics of exemplar retrieval in judgments from multiple cues.* [Conference presentation]. 15th Conference of the Section Methods and Evaluation in the German Psychological Society (DGPs), Mannheim, Germany.

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, *120*(1), 3–19. https://doi.org/10.1037/0096-3445.120.1.3

American Diabetes Association. (2015). Standards of Medical Care in Diabetes—2015 Abridged for Primary Care Providers. *Clinical Diabetes*, *33*(2), 97–111. https://doi.org/10.2337/diaclin.33.2.97

Anderson, J. R., & Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin & Review*, *8*(4), 629–647. https://doi.org/10.3758/BF03196200

Anderson, J. R., Reder, L. M., & Lebiere, C. (1996). Working memory: Activation limitations on retrieval. *Cognitive Psychology*, *30*(3), 221–256. https://doi.org/10.1006/cogp.1996.0007

APA. (2013). Science of psychology. https://www.apa.org/education-career/guide/science

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., Van Aken, M. A. G., Weber, H., & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119. https://doi.org/10.1002/per.1919

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

Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Sciences*, *9*(2), 83–89. https://doi.org/10.1016/j.tics.2004.12.003

Ashby, F. G., Waldron, E. M., Lee, W. W., & Berkman, A. (2001). Suboptimality in human categorization and identification. *Journal of Experimental Psychology: General*, *130*(1), 77–96. https://doi.org/10.1037/0096-3445.130.1.77

Batchelder, W. H., Colonius, H., Dzhafarov, E. N., & Myung, J. (Eds.). (2017). *New handbook of mathematical psychology* (Vol. 2). Cambridge University Press. https://doi.org/10.1017/9781139245913

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Bishop, F. K., Maahs, D. M., Spiegel, G., Owen, D., Klingensmith, G. J., Bortsov, A., Thomas, J., & Mayer-Davis, E. J. (2009). The carbohydrate counting in adolescents with Type 1 diabetes (CCAT) study. *Diabetes Spectrum*, *22*(1), 56–62. https://doi.org/10.2337/diaspect.22.1.56

Böhm, U., Marsman, M., Matzke, D., & Wagenmakers, E.-J. (2018). On the importance of avoiding shortcuts in applying cognitive models to hierarchical data. *Behavior Research Methods*, *50*(4), 1614–1631. https://doi.org/10.3758/s13428-018-1054-3

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Bott, L., & Heit, E. (2004). Nonmonotonic extrapolation in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 38–50. https://doi.org/10.1037/0278-7393.30.1.38

Bower, G. (1967). A multicomponent theory of the memory trace. *Psychology of Learning and Motivation*, *1*, 229–325. https://doi.org/10.1016/S0079-7421(08)60515-0

Brehmer, A. (1988). Grading as a quasi-rational judgment process. In J. Lowyck, C. Clark, & R. Halkes (Eds.), *Teacher thinking and professional action.* (pp. 129–137). Routledge.

Brehmer, A., & Brehmer, B. (1988). What have we learned about human judgment from thirty years of policy capturing? In B. Brehmer & C. Joyce (Eds.), *Human judgment: The SJT view* (pp. 75–114). North Holland. https://doi.org/10.1016/S0166-4115(08)62171-8

Brehmer, B. (1969). Cognitive dependence on additive and configural cue-criterion relations. *The American Journal of Psychology*, *82*(4), 490. https://doi.org/10.2307/1420442

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, *11*(1), 1–27. https://doi.org/10.1016/0030-5073(74)90002-6

Brehmer, B. (1994). The psychology of linear judgement models. *Acta Psychologica*, *87*, 137–154. https://doi.org/10.1016/0001-6918(94)90048-5

Brehmer, B., & Joyce, C. (1988). *Human judgment the SJT view*. North Holland. https://doi.org/10.1016/S0166-4115(08)X6086-8

Bröder, A. (2000). A methodological comment on behavioral decision research. *Psychologische Beiträge*, *42*, 645–662.

Bröder, A., & Gaissmaier, W. (2007). Sequential processing of cues in memory-based multiattribute decisions. *Psychonomic Bulletin & Review*, *14*(5), 895–900. https://doi.org/10.3758/BF03194118

Bröder, A., & Gräf, M. (2018). Retrieval from memory and cue complexity both trigger exemplar-based processes in judgment. *Journal of Cognitive Psychology*, *30*(4), 406–417. https://doi.org/10.1080/20445911.2018.1444613

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.

Bröder, A., & Hilbig, B. E. (2017). Urteilen und Entscheiden. In J. Müsseler & M. Rieger (Eds.), *Allgemeine Psychologie* (pp. 619–659). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-53898-8_17

Bröder, A., Newell, B. R., & Platzer, C. (2010). Cue integration vs. exemplar-based reasoning in multi-attribute decisions from memory: A matter of cue representation. *Judgment and Decision Making*, *5*(5), 326–338.

Bröder, A., & Undorf, M. (2019). Metamemory viewed through the judgment lens. *Acta Psychologica*, *197*, 153–165. https://doi.org/10.1016/j.actpsy.2019.04.011

Brooks, L. R., & Hannah, S. D. (2006). Instantiated features and the use of "rules." *Journal of Experimental Psychology: General*, *135*(2), 133–151. https://doi.org/https://doi.org/10.1037/0096-3445.135.2.133

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review*, *100*(3), 511–534. https://doi.org/10.1037/0033-295X.100.3.511

Brown, N. R., & Siegler, R. S. (1996). Long-term benefits of seeding the knowledge base. *Psychonomic Bulletin & Review*, *3*(3), 385–388. https://doi.org/10.3758/BF03210766

Brown, N. R., & Siegler, R. S. (2001). Seeds aren't anchors. *Memory & Cognition*, *29*(3), 405–412. https://doi.org/10.3758/BF03196391

Brunswik, E. (1952). *The conceptual framework of psychology.* University of Chicago Press.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*, 193–217. https://doi.org/10.1037/h0047470

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*, 432–459. https://doi.org/10.1037/0033-295x.100.3.432

Cooksey, R. W. (1996). The methodology of social judgement theory. *Thinking & Reasoning*, *2*(2-3), 141–174. https://doi.org/10.1080/135467896394483

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. https://doi.org/10.1037/0278-7393.23.4.968

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89. https://doi.org/10.1016/j.jmp.2015.10.003

Doherty, M. E., & Kurz, E. M. (1996). Social judgement theory. *Thinking & Reasoning*, *2*(2-3), 109–140. https://doi.org/10.1080/135467896394474

Earle, T. C., & Cvetkovich, G. (1988). Risk judgment, risk communication and conflict management. In B. Brehmer & C. Joyce (Eds.), *Human judgment the SJT view* (pp. 361–400). North Holland. https://doi.org/10.1016/S0166-4115(08)62179-2

Einhorn, H. J., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Linear regression and process-tracing models of judgment. *Psychological Review*, *86*(5), 465–485. https://doi.org/10.1037/0033-295X.86.5.465

Erdfelder, E. (1993). BINOMIX: A BASIC program for maximum likelihood analyses of finite and beta-binomial mixture distributions. *Behavior Research Methods, Instruments, & Computers*, *25*(3), 416–418. https://doi.org/10.3758/BF03204535

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107–140. https://doi.org/https://doi.org/10.1037/0096-3445.127.2.107

Estes, W. K. (1986). Memory storage and retrieval processes in category learning. *Journal of Experimental Psychology: General*, *115*(2), 155–174. https://doi.org/10.1037/0096-3445.115.2.155

Estes, W. K. (1994). *Classification and cognition*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195073355.001.0001

Farrell, S., & Ludwig, C. J. H. (2008). Bayesian and maximum likelihood estimation of hierarchical response time models. *Psychonomic Bulletin & Review*, *15*(6), 1209–1217. https://doi.org/10.3758/PBR.15.6.1209

Farrell, S., & Lewandowsky, S. (2010). Computational models as aids to better reasoning in psychology. *Current Directions in Psychological Science*, *19*(5), 329–335. https://doi.org/10.1177/0963721410386677

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press. https://doi.org/10.1017/CBO9781316272503

Fiedler, K. (1996). Expaining and simulating judgment bisases as an aggregation phenomenom in probabilistic, multiple-cue environments. *Psychological Review*, *103*(1), 193–214. https://doi.org/10.1037//0033-295x.103.1.193

Fischbein, E., Deri, M., Nello, M. S., & Marino, M. S. (1985). The role of implicit models in solving verbal problems in multiplication and division. *Journal for Research in Mathematics Education*, *16*(1), 3. https://doi.org/10.2307/748969

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, *31*(4), 271–288. https://doi.org/10.1080/1047840X.2020.1853461

Gaissmaier, W., Fific, M., & Rieskamp, J. (2011). Analyzing response times to understand decision processes. In M. Schulte-Mecklenbeck, A. Küberger, & R. Ranyard (Eds.), *A handbook of process tracing methods for decision research* (pp. 141–159). Psychology Press.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669. https://doi.org/10.1037/0033-295X.103.4.650

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. Oxford University Press.

Glöckner, A. (2009). Investigating intuitive and deliberate processes statistically: The multiple-measure maximum likelihood strategy classification method. *Judgment and Decision Making*, *4*(3), 15.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1166–1183. https://doi.org/10.1037/0278-7393.22.5.1166

Goldstein, W. M. (2004). Social judgment theory: Applying and extending Brunswik's probabilistic functionalism. In D. J. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making* (pp. 37–61). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470752937.ch3

Goldstein, W. M., & Hogarth, R. M. (1997). Judgment and decision research: Some historical context. *Research on judgment and decision making: Currents, connections, and controversies* (pp. 3–65). Cambridge University Press.

Gorman, C. D., Clover, W. H., & Doherty, M. E. (1978). Can we learn anything about interviewing real people from "interviews" of paper people? Two studies of the external validity of a paradigm. *Organizational Behavior and Human Performance*, *22*(2), 165–192. https://doi.org/10.1016/0030-5073(78)90011-9

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802. https://doi.org/10.1177/1745691620970585

Hahn, U., & Chater, N. (1998). Similarity and rules: Distinct? Exhaustive? Empirically distinguishable? *Cognition*, *65*(2-3), 197–230. https://doi.org/10.1016/S0010-0277(97)00044-9

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

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review*, *62*(4), 255–262. https://doi.org/10.1037/h0046845

Hammond, K. R., & Brehmer, B. (1973). Quasi-rationality and distrust. Implications for international conflict. In L. Rappoport & D. Summers (Eds.), *Human judgment and social interaction.* (pp. 338–391). Holt, Rinehart and Winston.

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

Hammond, K. R., Stewart, T. R., Brehmer, B., Steinmann, D. O., Kaplan, M., & Schwartz, S. (1975). Social judgment theory. *Human judgement and decision processes* (pp. 271–312). Academic Press. https://doi.org/10.1016/B978-0-12-397250-7.50016-7

Hammond, K. R., & Summers, D. A. (1965). Cognitive dependence on linear and nonlinear cues. *Psychological Review*, *72*(3), 215–224. https://doi.org/10.1037/h0021798

Hannah, S. D., & Brooks, L. R. (2009). Featuring familiarity: How a familiar feature instantiation influences categorization. *Canadian Journal of Experimental Psychology*, *63*(4), 263–275. https://doi.org/10.1037/a0017919

Haslbeck, J. M. B., Ryan, O., Robinaugh, D. J., Waldorp, L. J., & Borsboom, D. (2021). Modeling psychopathology: From data models to formal theories. *Psychological Methods*. https://doi.org/10.1037/met0000303

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*(2), 233–249. https://doi.org/10.1002/bdm.1958

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96–101. https://doi.org/10.3758/BF03202365

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*(4), 411–428. https://doi.org/10.1037/0033-295X.93.4.411

Hirschmüller, S., Egloff, B., Nestler, S., & Back, M. D. (2013). The dual lens model: A comprehensive framework for understanding self–other agreement of personality judgments at zero acquaintance. *Journal of Personality and Social Psychology*, *104*(2), 335–353. https://doi.org/10.1037/a0030383

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin*, *57*(2), 116–131. https://doi.org/10.1037/h0047807

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2013). Deliberation's blindsight: How cognitive load can improve judgments. *Psychological Science*, *24*(6), 869–879. https://doi.org/10.1177/0956797612463581

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2014). General pillars of judgment: How memory abilities affect performance in rule-based and exemplar-based judgments. *Journal of Experimental Psychology*, *143*(6), 2242–2261. https://doi.org/10.1037/a0037989

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. https://doi.org/10.1037/xlm0000241

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2019). Testing learning mechanisms of rule-based judgment. *Decision*, *6*(14), 305–334. https://doi.org/ https://doi.org/10.1037/dec0000109

Hoffmann, J. A., von Helversen, B., Weilbächer, R. A., & Rieskamp, J. (2018). Tracing the path of forgetting in rule abstraction and exemplar retrieval. *Quarterly Journal of Experimental Psychology*, *71*(11), 2261–2281. https://doi.org/10.1177/1747021817739861

Hout, M. C., Papesh, M. H., & Goldinger, S. D. (2013). Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(1), 93–103. https://doi.org/10.1002/wcs.1203

Izydorczyk, D., & Bröder, A. (2021). Exemplar-based judgment or direct recall: On a problematic procedure for estimating parameters in exemplar models of quantitative judgment. *Psychonomic Bulletin & Review*, *28*(5), 1–19. https://doi.org/10.3758/s13423-020-01861-1

Izydorczyk, D., & Bröder, A. (2022). What is the airspeed velocity of an unladen swallow? Modeling numerical judgments of realistic stimuli. *Manuscript submitted for publication.*

Izydorczyk, D., & Bröder, A. (in press). Measuring the mixture of rule-based and exemplar-based processes in judgment: A hierarchical bayesian approach. *Decision.*

Jamieson, R. K., Crump, M. J. C., & Hannah, S. D. (2012). An instance theory of associative learning. *Learning & Behavior*, *40*(1), 61–82. https://doi.org/10.3758/s13420-011-0046-2

Johansen, M., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology*, *45*(4), 482–553. https://doi.org/10.1016/S0010-0285(02)00505-4

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition*, *106*(1), 259–298. https://doi.org/10.1016/j.cognition.2007.02.003

Juslin, P., Olsson, H., & Olsson, A. C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology*, *132*(1), 133–156. https://doi.org/10.1037/0096-3445.132.1.133

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science*, *26*(5), 563–607. https://doi.org/10.1016/S0364-0213(02)00083-6

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of Linear Experts: Knowledge Partitioning and Function Learning. *Psychological Review*, *111*(4), 1072–1099. https://doi.org/10.1037/0033-295X.111.4.1072

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, *134*(3), 404–426. https://doi.org/10.1037/0033-2909.134.3.404

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. https://doi.org/10.3758/BF03193103

Karlsson, L., Juslin, P., & Olsson, H. (2008). Exemplar-based inference in multi-attribute decision making: Contingent, not automatic, strategy shifts? *Judgment and Decision Making*, *3*(3), 244–260.

Katahira, K. (2016). How hierarchical models improve point estimates of model parameters at the individual level. *Journal of Mathematical Psychology*, *73*, 37–58. https://doi.org/10.1016/j.jmp.2016.03.007

Kawamura, T., Takamura, C., Hirose, M., Hashimoto, T., Higashide, T., Kashihara, Y., Hashimura, K., & Shintaku, H. (2015). The factors affecting on estimation of carbohydrate content of meals in carbohydrate counting. *Clinical Pediatric Endocrinology*, *24*(4), 153–165. https://doi.org/10.1297/cpe.24.153

Klauer, K. C., & Kellen, D. (2018). RT-MPTs: Process models for response-time distributions based on multinomial processing trees with applications to recognition memory. *Journal of Mathematical Psychology*, *82*, 111–130. https://doi.org/10.1016/j.jmp.2017.12.003

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., ... Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

König, L. M., Ziesemer, K., & Renner, B. (2019). Quantifying actual and perceived inaccuracy when estimating the sugar, energy content and portion size of foods. *Nutrients*, *11*(10), 2425. https://doi.org/10.3390/nu11102425

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27. https://doi.org/10.1007/BF02289565

Lagnado, D. A., Newell, B. R., Kahan, S., & Shanks, D. R. (2006). Insight and strategy in multiple-cue learning. *Journal of Experimental Psychology: General*, *135*(2), 162–183. https://doi.org/10.1037/0096-3445.135.2.162

Lamberts, K. (2005). Mathematical modeling of cognition. In K. Lamberts & R. Goldstone (Eds.), *Handbook of cognition* (pp. 407–421). Sage Publication. https://doi.org/10.4135/9781848608177.n18

Laurenzi, A., Bolla, A. M., Panigoni, G., Doria, V., Uccellatore, A., Peretti, E., Saibene, A., Galimberti, G., Bosi, E., & Scavini, M. (2011). Effects of carbohydrate counting on glucose control and quality of life over 24 weeks in adult patients with type 1 diabetes on continuous subcutaneous insulin infusion. *Diabetes Care*, *34*(4), 823–827. https://doi.org/10.2337/dc10-1490

Lawson, R., & Bhagat, P. S. (2002). The role of price knowledge in consumer product knowledge structures. *Psychology & Marketing*, *19*(6), 551–568. https://doi.org/10.1002/mar.10024

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*(1), 1–15. https://doi.org/10.3758/PBR.15.1.1

Lee, M. D. (2018, March 23). Bayesian Methods in Cognitive Modeling. In J. T. Wixted (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience* (pp. 1–48). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119170174.epcn502

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332. https://doi.org/10.1037/0033-295X.111.2.309

Lykken, D. T. (1991). What's wrong with psychology anyway? In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology: Essays in honor of Paul E. Meehl* (pp. 3–39). University of Minnesota Press.

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*(4), 330–354. https://doi.org/10.1006/jesp.1998.1353

Mata, R., von Helversen, B., Karlsson, L., & Cüpper, L. (2012). Adult age differences in categorization and multiple-cue judgment. *Developmental Psychology*, *48*(4), 1188–1201. https://doi.org/10.1037/a0026084

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*(1), 24–42. https://doi.org/10.3758/BF03196347

McDaniel, M. A., Cahill, M. J., Robbins, M., & Wiener, C. (2014). Individual differences in learning and transfer: Stable tendencies for learning exemplars versus abstracting rules. *Journal of Experimental Psychology: General*, *143*(2), 668–693. https://doi.org/10.1037/a0032963

McElreath, R. (2020). *Statistical rethinking: A bayesian course with examples in R and Stan* (2nd ed.). CRC Press/Taylor & Francis Group. https://doi.org/10.1201/9780429029608

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Meehl, P. E. (1990a). Appraising and amending theories: The strategy of lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Meehl, P. E. (1990b). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195–244. https://doi.org/10.2466/pr0.1990.66.1.195

Mellers, B. A. (1980). Configurality in multiple-cue probability learning. *The American Journal of Psychology*, *93*(3), 429. https://doi.org/10.2307/1422722

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229. https://doi.org/10.1038/s41562-018-0522-1

Nilsson, H., Rieskamp, J., & Wagenmakers, E.-J. (2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology*, *55*(1), 84–93. https://doi.org/10.1016/j.jmp.2010.08.006

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058

Nosofsky, R. M. (1984). Choice, similarity and the context theory of classification. *Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104–114. https://doi.org/10.1037/0278-7393.10.1.104

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–61. https://doi.org/10.1037//0096-3445.115.1.39

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning,*

Memory, and Cognition, *14*(4), 700–708. https://doi.org/10.1037/0278-7393.14.4.700

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(1), 3–27. https://doi.org/10.1037/0096-1523.17.1.3

Nosofsky, R. M. (1992). Exemplar-based approach to relating categorization, identification, and recognition. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 363–393). Lawrence Erlbaum Associates, Inc.

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 18–39). Cambridge University Press. https://doi.org/10.1017/CBO9780511921322.002

Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(2), 282–304. https://doi.org/10.1037/0278-7393.15.2.282

Nosofsky, R. M., & Hu, M. (2022). Generalization in distant regions of a rule-described category space: A mixed exemplar and logical-rule-based account. *Computational Brain & Behavior*. https://doi.org/10.1007/s42113-022-00151-4

Nosofsky, R. M., & Little, D. R. (2010). Classification response times in probabilistic rule-based category structures: Contrasting exemplar-retrieval and decision-boundary models. *Memory & Cognition*, *38*(7), 916–927. https://doi.org/10.3758/MC.38.7.916

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(2), 266–300. https://doi.org/10.1037/0033-295X.104.2.266

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79. https://doi.org/10.1037/0033-295X.101.1.53

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

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*(6), 1371–1384. https://doi.org/10.1037/0278-7393.32.6.1371

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Pachur, T., & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology*, *65*(2), 207–240. https://doi.org/10.1016/j.cogpsych.2012.03.003

Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 534–552. https://doi.org/10.1037/0278-7393.14.3.534

Persson, M., & Rieskamp, J. (2009). Inferences from memory: Strategy- and exemplar-based judgment models compared. *Acta Psychologica*, *130*(1), 25–37. https://doi.org/10.1016/j.actpsy.2008.09.010

Platzer, C., & Bröder, A. (2013). When the rule is ruled out: Exemplars and rules in decisions from memory. *Journal of Behavioral Decision Making*, *26*, 429–441. https://doi.org/10.1002/bdm

Radev, S. T., D'Alessandro, M., Mertens, U. K., Voss, A., Köthe, U., & Bürkner, P.-C. (2020). Amortized bayesian model comparison with evidential deep learning. https://doi.org/10.48550/ARXIV.2004.10629

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356. https://doi.org/10.1111/1467-9280.00067

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. https://doi.org/10.1037/0096-3445.122.1.92

Richardson, D. C., & Spivey, M. J. (2000). Representation, space and Hollywood Squares: Looking at things that aren't there anymore. *Cognition*, *76*(3), 269–295. https://doi.org/10.1016/S0010-0277(00)00084-6

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*(2), 207–236. https://doi.org/10.1037/0096-3445.135.2.207

Rosner, A., & von Helversen, B. (2019). Memory shapes judgments: Tracing how memory biases judgments by inducing the retrieval of exemplars. *Cognition*, *190*, 165–169. https://doi.org/10.1016/j.cognition.2019.05.004

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*(2), 195–223. https://doi.org/10.3758/BF03257252

Rouder, J. N., Morey, R. D., & Pratte, M. S. (2018). Bayesian hierarchical models of cognition. In W. H. Batchelder, H. Colonius, E. N. Dzhafarov, & J. Myung (Eds.), *New handbook of mathematical psychology* (pp. 504–551). Cambridge University Press. https://doi.org/10.1017/9781139245913.010

Rouder, J. N., & Ratcliff, R. (2004). Comparing categorization models. *Journal of Experimental Psychology: General*, *133*(1), 63–82. https://doi.org/10.1037/0096-3445.133.1.63

Sacks, G., Veerman, J. L., Moodie, M., & Swinburn, B. (2011). 'Traffic-light' nutrition labelling and 'junk-food' tax: A modelled comparison of cost-effectiveness for obesity prevention. *International Journal of Obesity*, *35*(7), 1001–1009. https://doi.org/10.1038/ijo.2010.228

Schlegelmilch, R., Wills, A. J., & von Helversen, B. (2021). A cognitive category-learning model of rule abstraction, attention learning, and contextual modulation. *Psychological Review*. https://doi.org/10.1037/rev0000321

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325–345. https://doi.org/10.1007/BF02288967

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125–140. https://doi.org/10.1007/BF02289630

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323. https://doi.org/10.1126/science.3629243

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

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, *17*(4), 443–464. https://doi.org/10.3758/PBR.17.4.443

Shin, H. J., & Nosofsky, R. M. (1992). Similarity-scaling studies of dot-pattern classification and recognition. *Journal of Experimental Psychology: General*, *121*(3), 278–304. https://doi.org/10.1037/0096-3445.121.3.278

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational social psychology* (First, pp. 311–331). Routledge. https://doi.org/10.4324/9781315173726-14

Smaldino, P. E. (2019). Better methods can't make up for mediocre theory. *Nature*, *575*(7781), 9–9. https://doi.org/10.1038/d41586-019-03350-5

Smaldino, P. E. (2020). How to Build a Strong Theoretical Foundation. *Psychological Inquiry*, *31*(4), 297–301. https://doi.org/10.1080/1047840X.2020.1853463

Smith, E. R., & Zárate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review*, *99*(1), 3–21. https://doi.org/10.1037/0033-295X.99.1.3

Stewart, T. R., Moninger, W. R., Brady, R. H., Merrem, F. H., Stewart, T. R., & Grassia, J. (1989). Analysis of expert judgment in a hail forecasting experiment. *Weather and Forecasting*, *4*(1), 24–34. https://doi.org/10.1175/1520-0434(1989)004<0024:AOEJIA>2.0.CO;2

Stillesjö, S., Nyberg, L., & Wirebring, L. K. (2019). Building memory representations for exemplar-based judgment: A role for ventral precuneus. *Frontiers in Human Neuroscience*, *13*, 228. https://doi.org/10.3389/fnhum.2019.00228

Thibaut, J.-P., Gelaes, S., & Murphy, G. L. (2018). Does practice in category learning increase rule use or exemplar use—or both? *Memory & Cognition*, *46*(4), 530–543. https://doi.org/10.3758/s13421-017-0782-4

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. https://doi.org/10.1037/xlm0000696

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2013). Working memory capacity and retrieval from long-term memory: The role of controlled search. *Memory & Cognition*, *41*(2), 242–254. https://doi.org/10.3758/s13421-012-0261-x

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., Karlsson, L., Mata, R., & Wilke, A. (2013). Why does cue polarity information provide benefits in inference problems? The role of strategy selection and knowledge of cue importance. *Acta Psychologica*, *144*(1), 73–82. https://doi.org/10.1016/j.actpsy.2013.05.007

von Helversen, B., Karlsson, L., Rasch, B., & Rieskamp, J. (2014). Neural substrates of similarity and rule-based strategies in judgment. *Frontiers in Human Neuroscience*, *8*, 1–13. https://doi.org/10.3389/fnhum.2014.00809

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*(1), 220–229. https://doi.org/10.1037/a0016690

von Helversen, B., & Rieskamp, J. (2008). The mapping model: A cognitive theory of quantitative estimation. *Journal of Experimental Psychology*, *137*(1), 73–96. https://doi.org/10.1037/0096-3445.137.1.73

von Helversen, B., & Rieskamp, J. (2009a). Models of quantitative estimations: Rule-based and exemplar-based processes compared. *Journal of Experimental Psychology: Learning Memory and Cognition*, *35*(4), 867–889. https://doi.org/10.1037/a0015501

von Helversen, B., & Rieskamp, J. (2009b). Predicting sentencing for low-level crimes: Comparing models of human judgment. *Journal of Experimental Psychology: Applied*, *15*(4), 375–395. https://doi.org/10.1037/a0018024

Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, *39*(4), 202–217. https://doi.org/10.1037/teo0000137

Wills, A. J., & Pothos, E. M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin*, *138*(1), 102–125. https://doi.org/10.1037/a0025715

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, e49547. https://doi.org/10.7554/eLife.49547

Wirebring, L. K., Stillesjö, S., Eriksson, J., Juslin, P., & Nyberg, L. (2018). A similarity-based process for human judgment in the parietal cortex. *Frontiers in Human Neuroscience*, *12*, 1–18. https://doi.org/10.3389/fnhum.2018.00481

Wohldmann, E. L. (2013). Examining the relationship between knowing and doing: Training for improving food choices. *The American Journal of Psychology*, *126*(4), 449–458. https://doi.org/10.5406/amerjpsyc.126.4.0449

Wohldmann, E. L. (2015). Planting a Seed: Applications of cognitive principles for improving food choices. *The American Journal of Psychology*, *128*(2), 209. https://doi.org/10.5406/amerjpsyc.128.2.0209

Wohldmann, E. L., & Healy, A. F. (2020). Learning and transfer of calorie information. *Applied Cognitive Psychology*, *34*(6), 1485–1494. https://doi.org/10.1002/acp.3727

Zeigenfuse, M. D., & Lee, M. D. (2010). A general latent assignment approach for modeling psychological contaminants. *Journal of Mathematical Psychology*, *54*(4), 352–362. https://doi.org/10.1016/j.jmp.2010.04.001

A Acknowledgements

"*If I seem to wander, if I seem to stray, remember that true stories seldom take the straightest way.*"

from *The Name of the Wind*
by Patrick Rothfuss

Like true stories, this PhD journey was full of unexpected ups and downs, twists and turns. Luckily I had the support of many wonderful and brilliant people along the way, who made sure I did not get lost and that I could finish this journey.

First and foremost, I am deeply grateful to Arndt Bröder. Without him, this thesis would not exist. He fostered my interest for JDM research and brought my attention to the topics embedded in this thesis, thought me much about academic writing, careful experimentation, and scientific rigor. Regardless of how busy his schedule was, he always found time to answer my question, to provide feedback, or to give motivation and encouragement whenever it was needed. He gave me the freedom to pursue my own ideas and supported me in every way possible.

I am also grateful to Anna Baumert, who sparked my curiosity for psychological research and motivated me to pursue it myself. By always challenging me and giving me the freedom to do things in my own way, she made me learn many important things I used throughout my PhD.

I am thankful for the SMiP research training group for giving me the opportunity to do this PhD under the best circumstances possible. I thank Edgar Erdfelder, Benjamin Hilbig, Daniel Heck, and Jeff Rouder, who provided feedback, new insights, or advice at the many retreats, research colloquia, or anytime I needed it. I thank Annette Förster and Anke Söllner for their support throughout this PhD. Most importantly, I thank my fellow SMiPsters who created such a wonderful environment for a PhD, be it through the many stimulating talks during the retreats, or the after-workshop hours we enjoyed together.

I am especially thankful to my brilliant friends and colleagues who accompanied my on this journey and made it all the better. Malte, Stefan, Michi, Lili, Pascal, Sofia, and Tong, who always made me happy to go to work and their valuable advice on many occasions. Franziska and Nikoletta for their emotional support and

the countless coffee breaks and after work beers, which were often the highlight of the day. Many thanks to Martin, who made the office hours in B6 way more enjoyable, taught me much about hypothesis testing and Bayesian modeling, but more importantly, who always was ready to keep the SMiP traditions alive. Particular thanks to Sophie, who was like a second mentor to me and without whom I would not be here today. She was there for me from day one, answered all my questions about being a PhD and lecturer at the University of Mannheim, gave feedback to all my abstracts and manuscripts, and gave invaluable advice and moral support on every corner of my PhD.

I am also grateful for the wonderful group of friends, who walked with me long before this PhD journey. Leo, Kai, Saskia, Danielle, Frieda, Kristina, Simon, and Alex, for the countless happy moments we spend together and the reminder that there is a life beyond the PhD. Special thanks to David, who I could count on for 25 years and who was always there when I needed a friend. Although you often tried to keep me from working, your passion and drive were always an example for me.

However, this endeavor would not have been possible without my parents, Iris, Markus, Peter, and Ingeborg; my grandparents Klaus and Heidi; and my sister Sarah, who all unconditionally supported every decision I made in every way possible, enabled me to pursue my goals, and always encouraged me to do so. Most importantly, I am forever thankful to Judith for being my best friend and partner, for your unshakable believe in me, and for sharing every setback but also celebrating every success with me. Finally, I want to thank Karla. Although, you don't know it yet, you were for me the best reason both to work and not to work on this dissertation. You are the best thing that ever happened to me.

David Izydorczyk
Mannheim, September 2022

B Statement of Originality

1. I hereby declare that the presented doctoral dissertation with the title *Improving and Extending Models of Quantitative Judgments* is my own work.
2. I did not seek unauthorized assistance of a third party and I have employed no other sources or means except the ones listed. I clearly marked any quotations derived from the works of others.
3. I did not present this doctoral dissertation or parts of it at any other higher education institution in Germany or abroad.
4. I hereby conform the accuracy of the declaration above.
5. I am aware of the significance of this declaration and the legal consequences in case of untrue or incomplete statements.

I affirm in lieu of oath that the statements above are to the best of my knowledge true and complete.

Signature:

Date:

C Copies of Articles

© 2022, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final article will be available, upon publication, via its DOI: 10.1037/dec0000195

Measuring the mixture of rule-based and exemplar-based processes in judgment: A hierarchical Bayesian approach

David Izydorczyk & Arndt Bröder

University of Mannheim

Author Note

David Izydorczyk and Arndt Bröder, Department of Psychology, School of Social Sciences, University of Mannheim, Germany. Parts of this research were presented at the 15th conference of the DGPs Section Methods & Evaluation (Mannheim, Germany). This research was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG, GRK 2277) to the Research Training Group "Statistical Modeling in Psychology" (SMiP) and grant BR 2130/12-1 to the second author. The authors thank Sophie Scharf and Martin Schnuerch for helpful discussions and comments on an earlier version of the manuscript. The R scripts, results for all simulations and analyses, as well as the experimental data, RMarkdown file, with which this paper was written and which includes all code for analyses and figures, are available at the Open Science Framework (OSF, https://osf.io/7mabe/, Izydorczyk & Bröder, 2022, July 28)

Correspondence concerning this article should be addressed to David Izydorczyk, Experimental Psychology Lab, School of Social Sciences, University of Mannheim, D-68131 Mannheim, Germany. E-mail: izydorczyk@uni-mannheim.de

Abstract

Based on theoretical and empirical considerations, Bröder et al. (2017) proposed the RulEx-J model to quantify the relative contribution of rule- and exemplar-based processes in numerical judgments. In their original paper, a least-squares optimization procedure was used to estimate the model parameters. Despite general evidence for the validity of the model, the authors suggested that a strong bias in favoring the rule module could arise when there is noise in the data. In this article, we present a hierarchical Bayesian implementation of the RulEx-J model with the goal to rectify this problem. In a series of simulation studies, we demonstrate the ability of the hierarchical Bayesian RulEx-J model to recover parameters accurately and to be more robust against noise in the data, compared to a least-squares estimation routine. One further advantage of the hierarchical Bayesian approach is the direct implementation of hypotheses about group differences in the model structure. A validation experiment as well as reanalyses of two experiments from different labs demonstrate the usefulness of the approach for testing hypotheses about processing differences. Further applications for judgment research are discussed.

Keywords: numerical judgments, rule-based processes, exemplar-based processes, hierarchical Bayesian modeling

Word count: 10805

Measuring the mixture of rule-based and exemplar-based processes in judgment: A hierarchical Bayesian approach

Introduction

Every day, we have to make numerous judgments about continuous variables, such as the calorie content of a dessert, the dangerousness of crossing a busy street or the temperature outside. If the judgment is expressed on a numerical scale, it is termed a quantitative judgment. At least two different types of processes have been proposed to account for quantitative judgments: *Rule*-based and *exemplar*-based processes (Brehmer, 1994; Einhorn et al., 1979; Juslin et al., 2003; Karlsson et al., 2008; von Helversen & Rieskamp, 2009). Based on empirical evidence and methodological considerations, Bröder et al. (2017) proposed the *RulEx-J* model, which assumes assumes that both processes work in parallel and that the final judgment is a mixture of both distinct processes. The goal of this article is to introduce and test a hierarchical Bayesian implementation of the RulEx-J model, which improves upon the original parameter estimation method (Bröder et al., 2017). The remainder of this article is structured as follows: We first give a short summary about rule- and exemplar-based processes and how they interact, as well as problems with the original RulEx-J model. We then formally introduce the RulEx-J model and discuss problems with its current implementation in more detail. Next, we present the hierarchical Bayesian implementation of the RulEx-J model as a way to improve upon these problems. We then present a series of simulations that examine the ability of the model to recover parameters and the robustness against different magnitudes of noise in the data. Furthermore, we apply the hierarchical Bayesian model to data of a new experiment, aimed at validating the process mixing parameter α of the RulEx-J model (for more details, see below). We also reanalyse two existing data sets of experiments, using different manipulations and stimuli, to check whether previous results can be reproduced. [1]

[1] All R scripts, the JAGS model codes and result files are available at the Open Science Framework of this project (https://osf.io/7mabe/). All simulations and analyses were conducted using R (Version 4.2.0; R Core Team, 2020) and the R-packages *doSNOW* (Version 1.0.20; Corporation & Weston, 2019), *dplyr* (Version 1.0.9; Wickham et al., 2020), *foreach* (Version 1.5.2; Microsoft & Weston, 2020), *ggplot2* (Version 3.3.6; Wickham, 2016), *knitr* (Version 1.39; Xie, 2015), *papaja* (Version 0.1.0.9999; Aust & Barth, 2020), *polspline*

Processes of quantitative judgments and how they interact

Based on Brunswik's lens model (Brunswik, 1955), researchers assume that in rule-based processing people combine and integrate cue information according to a learned rule (Hoffmann et al., 2019). This could, for instance, be a weighted linear additive rule (e.g., Brehmer, 1994; Juslin et al., 2003) or a simpler heuristic, which ignores part of the cue information. For example, the cues "sweetness", "estimated amount of cream", and "size" of a dessert might form the basis for additively combining them into an estimate of its calorie content. By contrast, exemplar-based processes are not based on the abstraction and learning of cue-criterion relations. Rather, exemplar-based processes assume that people store previously encountered objects and their criterion values in long-term memory (Juslin et al., 2003, 2008). New objects are then judged based on the similarity to the exemplars stored in memory (Juslin et al., 2003; Medin & Schaffer, 1978; Nosofsky, 1984). For instance, judging the calorie content of a dessert might be based on the similarity to past desserts, of which the calorie content was known. The models describing exemplar-based processes have originated in the domains of memory (e.g., Hintzman, 1984) as well as categorization and classification (e.g., Medin & Schaffer, 1978; Nosofsky, 1984). However, sparked by the important work of Juslin and colleagues (e.g., Juslin & Persson, 2002; Juslin et al., 2003), the application and impact of exemplar models in the areas of judgment and decision making has increased during the last two decades (e.g., Bröder & Gräf, 2018; Hoffmann et al., 2013; Juslin et al., 2003; Mata et al., 2012; Pachur & Olsson, 2012; Persson & Rieskamp, 2009; von Helversen & Rieskamp, 2009).

Initially, researchers proposed a division of labor between both, rule-based and exemplar-based processes, where individuals would use only one process at a time across all trials (or at least within trials), but would shift between these qualitatively different processes, contingent on the structure of the task (e.g., Juslin et al., 2003, 2008; Karlsson et al., 2008; Pachur & Olsson, 2012; von Helversen et al., 2010). In their thorough individual differences analysis, Hoffmann et al.

(Version 1.1.20; Kooperberg, 2020), *Rcpp* (Version 1.0.8.3; Eddelbuettel & Balamuta, 2017; Eddelbuettel & François, 2011), *runjags* (Version 2.2.1.7; Denwood, 2016), *tibble* (Version 3.1.7; Müller & Wickham, 2020), and *truncnorm* (Version 1.0.8; Mersmann et al., 2018). The Bayesian models were implemented with JAGS (Plummer, 2003) Version 4.3.0.

(2014) validated the distinction between both processes by showing that they draw on different cognitive resources. According to their analysis, rule-based processing relies on working memory whereas exemplar-based processing rather depends on long-term memory. Using functional magnetic resonance imaging (fMRI), von Helversen, Karlsson, et al. (2014) found that rule-based and exemplar-based processes involve different neural correlates and different patterns of neural activation (cf., Wirebring et al., 2018). The methods used to measure the use of rule-based and exemplar-based processing in a given task condition reflected this dichotomous characterization of the judgment process. For instance, researchers would classify participants as users of a rule- or exemplar-based strategy (reflecting the corresponding cognitive process) based on the best-fitting model (e.g., Bröder et al., 2010; Pachur & Olsson, 2012; Persson & Rieskamp, 2009; Platzer & Bröder, 2012).

As an alternative to assuming a shift between qualitatively different processes, recent research suggests that there might be a "blending" or a mixture of both processes (e.g., Albrecht et al., 2019; Bröder et al., 2017; Herzog & von Helversen, 2018; Hoffmann et al., 2014; von Helversen, Herzog, & Rieskamp, 2014; Wirebring et al., 2018). For example, von Helversen, Herzog, and Rieskamp (2014) had their participants learn to judge the suitability of six training employees on a scale from 0 to 100. The job suitability was determined by a simple linear additive rule based on four cues (quality of work experience, motivation, skills, and education). Results showed that the judgments of new employees where influenced by the facial similarity to previously encountered exemplars, even though participants had all information to use the simple learned rule and using facial similarity led to worse judgments than ignoring it. These results are in line with other empirical evidence which suggests that exemplar retrieval and rule knowledge interact in category or continuous judgments. For example, Erickson and Kruschke (1998) showed that although participants were able to use a learned rule to categorize new stimuli, the similarity of specific training exemplars still affected classification probabilities. In addition, research by Brooks and colleagues (Allen & Brooks, 1991; Brooks & Hannah, 2006; Hannah & Brooks, 2009; Regehr & Brooks, 1993) showed that the similarity of features or exemplars affected classification speed or accuracy, even when a perfectly predictive classification rule was present and sometimes even explicitly given to the participants. Building up on these experiments, Hahn et al. (2010) found

similarity effects on accuracy or response times even though the manipulated similarity was irrelevant to the category membership and there were very simple, explicit, and perfectly predictive three- (Exp. 1, 3, & 4) or one-feature (Exp. 2) rules availabe. Their suggestion, that the influence of similarity is probably automatic and beyond strategic control is in line with findings from Macrae et al. (1998), who showed that automatic and unintentionally activated exemplars can lead to a decrease in performance even in simple tasks. Wirebring et al. (2018) found that brain activations associated with exemplar-based judgment processes where apparent even in conditions where the behavioral response was guided by a rule-based strategy. Finally, Herzog and von Helversen (2018) argue that from a mere normative and ecological perspective a mixture of processes can lead to more accurate judgments than relying on a single strategy.

The coarse-grained analysis of classifying participants as users of either a rule- or an exemplar-based strategy cannot detect subtle mixes of both processes as suggested by these studies. Therefore, based on these empirical findings and methodological considerations, Bröder et al. (2017) proposed the RulEx-J model as a measurement model to estimate the relative contribution of rule-based and exemplar-based processing in quantitative judgments. This model incorporates the idea of a process mix in cue-based judgments in line with former research (e.g., Hahn et al., 2010; von Helversen, Herzog, & Rieskamp, 2014; Wirebring et al., 2018).

The RulEx-J model in Bröder et al (2017)

Up to now the parameters of the RulEx-J model and similar blending models were estimated by using maximum-likelihood (ML) or least-squares (LS) optimization procedures (e.g., Albrecht et al., 2019; Bröder & Gräf, 2018; Bröder et al., 2017). In the article presenting the RulEx-J model, using these parameter-estimation approaches, Bröder et al. (2017) suggested a strong bias in favoring the rule module when the data became noisier. This is because the rule module is more complex than the exemplar module and thus able to fit the noise in the data better. This behavior of favouring the rule module is a strong disadvantage, since many researchers are interested in what aspects of the environment, learning phase, or judgment task influence the predominant type of processing (e.g., Bröder et al., 2010; Juslin et al., 2003, 2008; Karlsson et al., 2008; Pachur & Olsson, 2012; Trippas & Pachur, 2019; von Helversen et al., 2010). An artificial bias

towards rule-based processing might thus lead to wrong conclusions. For example, an experimental manipulation could affect the reliability of a cognitive process (by increasing random noise) without affecting its nature. Still, this would show as a processing difference in the original RulEx-J model. A more promising way of estimating the model parameters and thus the relative contribution of each process is a hierarchical Bayesian approach.

In the next sections, we introduce the RulEx-J model and discuss problems with its current implementation in more detail. We then present a hierarchical Bayesian implementation of the RulEx-J model as a way to improve upon these problems.

RulEx-J

The RulEx-J model is foremost intended as a measurement model to determine the relative contribution of rule- and exemplar-based processes in people's numerical judgments (Bröder et al., 2017). Instead of assuming that participants use either a rule- or an exemplar-based processes to make their judgments, the RulEx-J model assumes that both processes work in parallel and that the final judgment is a mixture of both distinct processes. Hence, people's judgments are conceptualized as a blending of rule- and exemplar-based processes. Similar to the ATRIUM model (Erickson & Kruschke, 1998), when a probe is presented to a person, it will be processed by an exemplar module E and a rule module R, each making their distinct tentative judgments. According to the RulEx-J model, the actual final judgment J is a weighted combination of both interim judgments:

$$J = \alpha J_R + (1 - \alpha) J_E, \tag{1}$$

where α is the mixture parameter, and J_R and J_E are the judgment outputs from the respective rule or exemplar module[2]. The α parameter is the main parameter of interest of the model and this article, since it measures the relative impact of rule- and exemplar-based processes on the final judgment. The α parameter can range from 0 to 1, with larger values indicating more rule-based

[2] This implementation of a mixture between processes assumes that both processes work independently and in parallel and is only one possible implementation of a mixture process (for more see Section Limitations and future directions).

processing and smaller values indicate more exemplar-based processing. However, the estimate of α will depend on the actual set of stimuli used for estimation, since, different sets of exemplars, cue patterns, and criterion values will differ in their ability to differentiate between the processes. Thus, instead of interpreting the absolute α values, one should compare the α values across experimental conditions using stimuli of similar logical structure (Bröder et al., 2017).

In the next sections, we first introduce the formal models which are used to model the rule- and exemplar-based processes in the respective module. Subsequently, we introduce the hierarchical Bayesian implementation of the RulEx-J model which we use throughout the rest of this article.

The rule module

The rule module is implemented as a linear regression model (Einhorn et al., 1979; Juslin et al., 2008). The judgment J_R of a probe $\vec{p}$ with n binary cues is generated by

$$J_R = w_0 + \sum_{j=1}^{n} \text{cue}_j w_j, \tag{2}$$

where w_0 is an intercept and w_j, for $j \neq 0$, are the cue weights, which can be interpreted as cue utilizations. This linear combination framework is quite flexible and can mimic simpler strategies focusing on one or only a few cues by choosing appropriate (zero) cue weights.

The exemplar module

The exemplar module is represented by the *context model* (Medin & Schaffer, 1978) extended to numerical judgments (see, Juslin & Persson, 2002). The model is based on the similarity S between a probe and the exemplars. It is assumed that the probe serves as a retrieval cue, activating previously encountered exemplars in memory. The probe $\vec{p}$ and each exemplar $\vec{e}$ are again represented by vectors of n binary cues. The similarity parameters s_j, $j = 0, ..., n$ are the only free parameters in this model, defined on the interval $(0, 1]$. They determine how strongly a mismatch on cue j between probe and exemplar influences the perceived similarity between probe and exemplar that can vary between (almost) 0 and 1. For simplicity, we assume the s_j to be constant across cues, that is, $s_j = s$, (e.g., Bröder & Gräf, 2018; Juslin & Persson, 2002; von

Helversen & Rieskamp, 2008)[3]. The similarity $S(\vec{p}, \vec{e}_k)$ between probe $\vec{p}$ and one exemplar $\vec{e}_k$ is determined according to the similarity rule of the context model (Medin & Schaffer, 1978):

$$S(\vec{p}, \vec{e}) = \prod_{j=1}^{n} d_j \text{ with } d_j = \begin{cases} 1 \text{ if } p_j = e_j \\ s \text{ if } p_j \neq e_j \end{cases} \tag{3}$$

where n is the number of cues of each object. For binary cues and assuming the same s-parameters for all features this simplifies to:

$$S(\vec{p}, \vec{e}) = s^{n-m}, \tag{4}$$

where m is the number of matching cues between $\vec{p}$ and $\vec{e}_k$. The judged criterion value J_E of the probe $\vec{p}$ is then the average of all n_c exemplar criterion values c in memory, weighted by the similarity of the respective exemplar to the probe:

$$J_E = \frac{\sum_{k=1}^{n} S(\vec{p}, \vec{e_k}) c(\vec{e_k})}{\sum_{k=1}^{n} S(\vec{p}, \vec{e_k})}, \tag{5}$$

where $c(\vec{e_k})$ is the criterion value of exemplar k.

Problems with the RulEx-J model and advantages of a Bayesian hierarchical solution

In this paper, we introduce a hierarchical Bayesian version of the RulEx-J model since the hierarchical Bayesian modeling framework offers many advantages and has therefore become a very popular tool for estimating latent parameters of cognitive models (e.g., Bott et al., 2020; Mattes et al., 2020; Schlegelmilch & von Helversen, 2020; Schubert et al., 2019; for general introductions see Lee, 2018; McElreath, 2020; Rouder et al., 2018). For instance, the hierarchical structure of the model naturally reflects the hierarchical data structure of many experiments, where several

[3] There are also empirical data showing that this simplified version outperforms the more complex model with a separate s_j parameter for each cue j in predicting individuals behavior (von Helversen & Rieskamp, 2008, 2009).

participants perform multiple trials of the same task and it is the aim of the researcher to draw conclusions on the group level (e.g., Steingroever et al., 2018). Instead of assuming that all individuals are the same (i.e., complete pooling approach) or that there are no informative similarities between individuals (i.e., no pooling approach) , hierarchical models assume that there is some similarity between individuals and, thus, they use the information from each individual to inform the estimates of other individuals, while taking into account that some participants might allow for more informative and reliable estimates than others (Gelman et al., 2014; McElreath, 2020). It has been shown that this partial pooling of information can lead to more accurate estimates (Efron & Morris, 1977; Farrell & Ludwig, 2008; Katahira, 2016; Rouder & Lu, 2005; Rouder et al., 2007)[4]. The reason is that individual parameters can be described by a group-level distribution which, given by the hierarchical structure, allows individual estimates to be informed by other individuals in a sample. Individual parameter estimates that are deemed unlikely given the overall group-level distribution of parameter values (because they are located at the extremes of the distribution) or are unreliable (because they have a large uncertainty) are pulled closer towards the group mean. This property called *shrinkage* is a result from regularization and leads to less overfit and more accurate estimates on average, than when parameters are estimated separately on an individual level (Gelman et al., 2014; McElreath, 2020). For these reasons, it has been argued that hierarchical methods provide a more thorough and efficient evaluation of models in cognitive science (Rouder et al., 2005; Shiffrin et al., 2008; van Ravenzwaaij et al., 2011). The pooling of information of hierarchical Bayesian models is especially useful when there is only a limited number of data available for each individual (Katahira, 2016; McElreath, 2020), as is common in many multiple-cue judgment studies. Since these studies rely on the learning of exemplars and cues, the number of trials of each person is often small. For instance, in a non-exhaustive literature search, the median number of stimuli in the judgment phase was 16, ranging from 9 to 100 (see the supplement file in the online materials). Although hierarchical models are not exclusive to the

[4] However, the hierarchical structure is an assumption of the model about individual differences and how latent parameters of participants are related to each other. Thus, hierarchical models can also lead to less accurate estimates in some cases, when the hierarchical assumptions deviate from the underlying properties of the data (Scheibehenne & Pachur, 2015)

Bayesian modeling framework, its flexibility makes it easy to implement hierarchical structures for more complex cognitive models.

A hierarchical Bayesian approach not only can increase the accuracy of parameter estimates of individuals, but also allows to make better inferences about group differences. Boehm et al. (2018) showed that the common two-step approach, where parameters are estimated separately for each individual and then subsequent tests (e.g., *t*-test, ANOVA) are performed on these individual parameters, can lead to biased inferences. In comparison, the flexibility of the Bayesian modeling framework allows to directly model group differences of latent parameters (Boehm et al., 2018).

Furthermore, as suggested by Bröder et al. (2017), one problem with their parameter estimation method (LS) is that the RulEx-J model strongly favors a rule-based processing when there is substantial noise in the data. The parameter estimates of α will tend to be biased towards 1.0, since the rule module has more free weight parameters (e.g., five when there are four cues) than the exemplar module, which has only one parameter per participant[5] (the s parameter), and thus is more able to (over)fit the noise in the data[6]. We assume that a Bayesian approach will reduce this bias, since the different complexity of the exemplar- and rule-modules are automatically taken into account.

Therefore, by using a hierarchical Bayesian modeling approach, we aim to improve on the shortcomings and problems of the original parameter-estimation method used by Bröder et al. (2017) and present interested researchers with a tested and state-of-the-art alternative.

---

[5] This difference in number of parameters is partially due to the choice of making equality constraints for the parameters in the exemplar module, where the s_i parameter of each cue i are constrained to be the same value. Without this constraint, the exemplar model would have only one parameter less than the rule model. See the section *The exemplar module* above

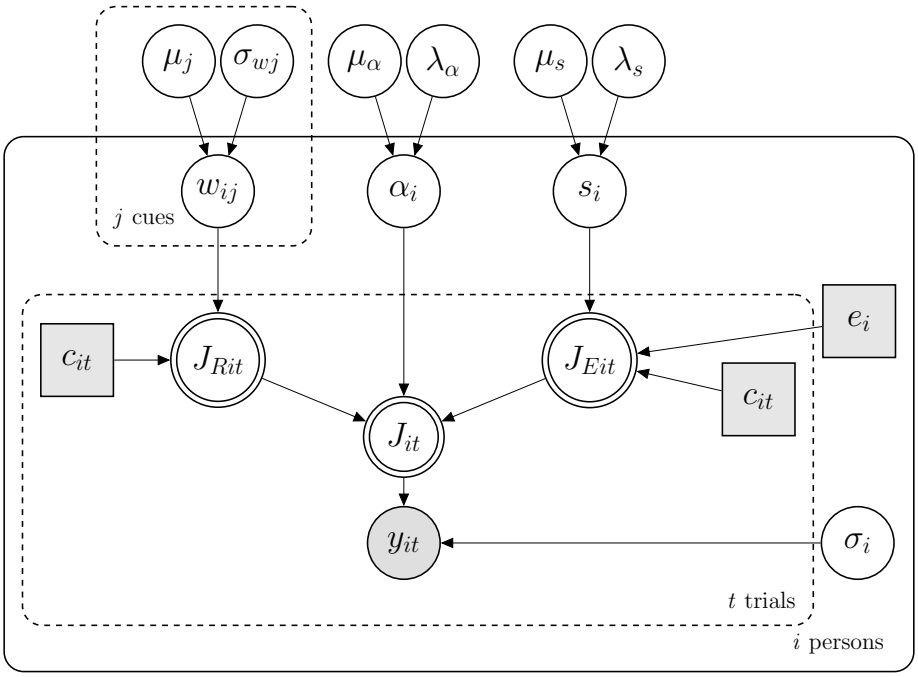
[6] The number of parameters of a model is only one factor determining the complexity of the model. Other factors such as the parameter range and the functional form (i.e., how the parameters are combined) also influence a model's complexity.

The hierarchical Bayesian RulEx-J model

The graphical model of the hierarchical Bayesian RulEx-J model is depicted in Figure 1. We use the notation of Lee (2008), in which observed variables (i.e., the data) are shown as shaded nodes and unobserved variables (i.e., model parameters to be inferred) are shown as unshaded nodes. Discrete variables are indicated by square nodes and continuous variables are indicated by circular nodes. Stochastic variables are indicated by single-bordered nodes, and deterministic variables are indicated by double-bordered nodes.

Figure 1

Graphical model representation of the hierarchical Bayesian RulEx-J model.



Like the original RulEx-J model, the Bayesian hierarchical version assumes that the response y_{it} of the ith participant in a given trial t is based on a weighted average of a rule-based and an exemplar-based process.

For the rule module, the weight parameter w_{ij} of the ith person and jth cue is assumed to be normally distributed with a corresponding mean μ_j and a general standard deviation σ_w[7]. Thus,

[7] Note, that in JAGS the normal distribution is parameterized in terms of precision τ and not standard

we assume that for each specific cue the weight of a person is randomly distributed around a cue specific mean (μ_j). The predicted judgment of the rule module J_{Rit} of the ith person in the tth trial is then computed based on Equation 2 and the corresponding cues c_{it} of the stimulus in this trial and of this person.

For the exemplar module, the individual s parameters are drawn from a group-level Beta(μ_s,λ_s) distribution, defined on the interval (0,1] to reflect the boundaries of the s parameter[8]. The group-level hyperparameters μ_s and λ_s are not the standard shape parameters of the Beta distribution (i.e., a_s and b_s). Rather μ_s and λ_s can be conceived as the mean and a measure of precision of the group-level distributions and thus, can be more meaningfully interpreted than the a_s and b_s parameters (Ferrari & Cribari-Neto, 2004; Lee & Wagenmakers, 2013). The a_s and b_s shape parameters from the Beta distribution can then be computed from μ_s and λ_s via $a_s = \mu_s \times \lambda_s$ and $b_s = (1 - \mu_s) \times \lambda_s$. The predicted judgment of the exemplar module J_{Eit} of each person i in each trial t is then computed based on Equations 4 and 5 and the corresponding cues c_{it} of the stimulus in this trial and of this person, as well as the exemplars e_i learned by the respective person i.

Like the s_i parameters, we assumed that the α_i parameters of each person i follow a group-level Beta(μ_α,λ_α) distribution. The final predicted judgment J_{it} of each person i in each trial t is then computed according to Equation 1.

The observed judgment y_{it} of the ith participant in the tth trial is given by a normal distribution centered around the final predicted judgment J_{it} with some precision σ_i.

Simulations

In this section, we present the results of two simulation studies. In the first simulation, we assessed whether the hierarchical Bayesian implementation of the RulEx-J model could accurately recover parameter values, which is necessary if we want to apply the model to real data, where the

---

deviation σ or variance σ^2. In the model code, we therefore transform the standard deviations to precision with $\tau = \frac{1}{\sigma^2}$.

[8] In the model, we used lower and upper bounds of 0.001 and 0.999 to avoid possible problems on the parameter boundaries.

true values of the parameters are not known. In the second simulation, we assessed the robustness and behavior of the hierarchical Bayesian RulEx-J model when there is noise in the data. These conditions are more similar to empirical data and thus might reveal certain caveats when applying the Bayesian hierarchical RulEx-J model. To test the robustness of the hierarchical Bayesian RulEx-J model against noise in the data, we generated judgment data with various levels of noise and with different underlying similarities between the α parameters of the synthetic participants. We then estimated the parameters using the hierarchical Bayesian RulEx-J model and with the least-squares optimization routine as in the original paper (Bröder et al., 2017). We suspected that the hierarchical Bayesian model would be more robust against error than both non-hierarchical versions. We also expected that the hierarchical Bayesian RulEx-J model would more accurately recover the α parameters of different synthetic participants, the more similar the individual true parameters were. Although we report the results for all individual-level parameters (α, s, w_j), the α parameter is the parameter of central interest and of major relevance for the questions in this line of research. In the following sections, we first present how we generated the simulated data and how parameters were estimated, before presenting the results.

Method

Data generation

In the first step of the simulations, we generated a stimulus matrix, consisting of 32 stimuli that can be created with five binary cues. The criterion values of the stimuli were computed according to a linear additive rule:

$$c = w_{0_{gen}} + \text{cue}_1 w_{1_{\text{gen}}} + \text{cue}_2 w_{2_{\text{gen}}} + \text{cue}_3 w_{3_{\text{gen}}} + \text{cue}_4 w_{4_{\text{gen}}} + \text{cue}_5 w_{5_{\text{gen}}} \tag{6}$$

where cue_j represents the binary cues coded with 0 and 1 and $w_{j_{\text{gen}}}$ the corresponding cue weights used for generating the criterion values. Of these 32 stimuli, 16 were randomly selected as exemplars. To avoid a perfect linear predictability of the criterion and, thus, to make the predictions of the rule and exemplar model differentiable (Bröder & Gräf, 2018; Bröder et al., 2017), the eight most extreme stimuli (i.e., the four stimuli with the highest and the four stimuli with the lowest criterion

value) were never selected as exemplars. We also switched the criterion values between three pairs of exemplars, that is, if one exemplar a of this switch pair would have a criterion value of 31 and exemplar b of the pair a value of 59, the new values after switching would be 59 for a and 31 for b. The cue weights $w_{j_{\text{gen}}}$ for cues $j = 0, ..., 5$ had to sum to 100. For cues $j = 1., ..., 5$ the weights were randomly drawn from a truncated normal distribution with $\mu = 15$, $\sigma = 10$, an upper bound of 100, and a lower bound of 1. The value of the intercept $w_{0_{\text{gen}}}$ was drawn from a truncated normal distribution with $\mu = 10$, $\sigma = 1$, an upper bound of 100, and a lower bound of 1.

In the second step, we drew the generating parameter values for $n = 30$ simulated participants in the first simulation, which is a typical sample size in such experiments (e.g., Bröder et al., 2017; Hoffmann et al., 2013; Trippas & Pachur, 2019)), and $n = 50$ in the second simulation. In the first simulation, the α parameter values were drawn from a uniform Beta(1,1) distribution and in the second simulation from a uniform Beta(1,1), a Beta(5,5), or peaked Beta(15,15) distribution, simulating different levels of underlying similarities between participants (see Figure 2 for an illustration of the resulting distributions). The s parameter values were drawn from a slightly skewed Beta(3,5) distribution which reflects a sensible range of s parameter values found in experimental studies (Izydorczyk & Bröder, 2021). The parameter values for the cue weights w_j were drawn from a truncated normal distribution with $\mu = w_{j_{\text{gen}}}$, $\sigma = 1$, an upper bound of 100, and a lower bound of 1. Thus, the parameter values of the cue weights w_j of the rule module of each participant were distributed around the corresponding cue weight $w_{j_{\text{gen}}}$ which was used to generate the criterion values of the stimuli. This reflects the idea of participants learning the cue weights in an experiment.

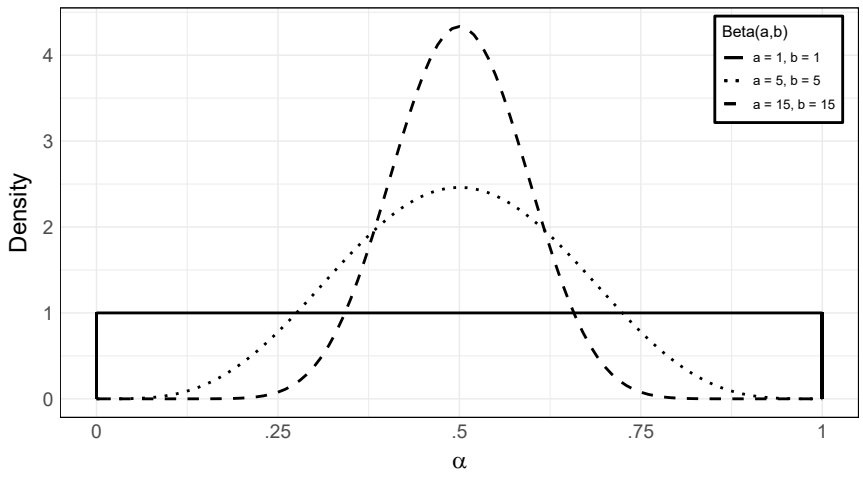
In the third step, judgment data for each simulated participant were generated with the RulEx-J model according to the drawn parameter values of Step 2 and the generated stimulus matrix in Step 1. In the second simulation, we added normal distributed error to the generated judgments of each person with $\mu = 0$ and $\sigma_\epsilon = 0$, 2, 4 or 8. We then estimated the parameters with the Bayesian RulEx-J model in both simulations, and also using LS-estimation in the second simulation. Next, we computed the root-mean-squared-error (RMSE) as a measure of absolute deviation of the estimated posterior mean of each parameter from the corresponding true parameter

values as a measure of parameter recovery accuracy in both simulations.

All steps were repeated 100 times in the first simulation. Since there were 12 different simulation design combinations in the second simulation, we reduced the number of repetitions from 100 to 50 in order to reduce the time needed to run the simulation. Parallelizing the repetition over 60 cores still took the simulation 80h to complete. Given the reduced number of repetitions, we increased the number of simulated participants from $n = 30$ to $n = 50$ in order to reach a similar overall sample size as in the first simulation.

Figure 2

Illustration of the Beta distribution for different values of the shape parameters a and b



Prior distributions

Based on the way we generate our simulated data and the underlying true parameters as described in the previous section, we used a Normal($\mu = 20$, $\sigma = 40$) prior for the group-mean parameters μ_j. We also set a lower bound of 0 and an upper bound of 100 on μ_j based on the possible range of values in our simulation. This truncated normal prior corresponds to giving the most weight to simulation specific sensible values, while still having a large amount of uncertainty.[9] For the group-level cue-weights standard deviation σ_w we used a weak Exponential(0.5) prior, which gives more weight to smaller values, indicating more similarity of the cue weights between

[9] We also tested the model with a $\mu_j \sim$ Uniform(0,100) prior. Since results of this simulation do not differ from those reported here, we stayed with the more informative and reasonable $\mu_j \sim$ Normal(20, 40) prior.

participants. For the group-level parameters μ_s and λ_s, we chose priors of $\mu_s \sim$ Beta(1,1) and $\lambda_s \sim$ Uniform(1,100), so that the resulting prior distribution of the subsequent individual s_i parameters was uniform. We used the same priors for μ_α and λ_α. Finally, we used again a weak Exponential(0.5) prior for σ_i.

Parameter estimation

In both simulations, the posterior distributions of the parameters were estimated by using Markov Chain Monte Carlo (MCMC) sampling. All of our simulation results are based on MCMC chains with 10,000 samples from each of two independent chains[10], collected after 20,000 burn-in samples were discarded, 20,000 adaptive iterations, and thinning by recording every 35th sample. Convergence of the MCMC chains was assessed for one iteration of the simulation by visual inspection and the $\hat{R}$ statistic ($\hat{R} \leq 1.02$ for all parameters, Gelman and Rubin (1992), see the example of the MCMC traces in the online materials, referred to in Footnote 1). We then used the means of the posterior distributions as estimates of the respective parameters.

Simulation Results

How well does the model recover parameters?

We found very good parameter recovery for the α (RMSE = 0.01), s (RMSE = 0.02), and cue weight parameters (RMSE $\leq$ 0.27) over all repetitions of the first simulation, as indicated by the low RMSE values. The intercept parameter w_0 showed the worst parameter recovery results (RMSE = 0.54), see also Figures 3, 4, and 5.

How does the model behave when there is noise in the data

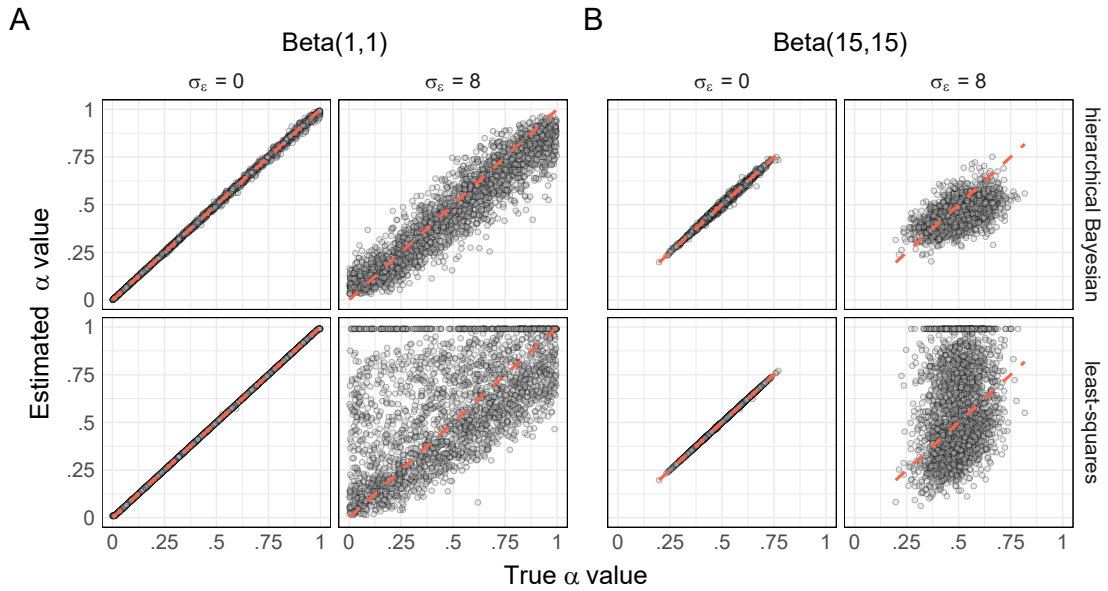
The results for the second simulations with the largest amount of noise ($\sigma_\epsilon = 8$) are shown in Table 1. The full results can be found in the online materials of this project.

[10] We used only two chains here to reduce the computation time and demand of the simulation. However, we checked the convergence in one run of the simulation beforehand using three chains and we would recommend using more than two chains in actual applications.

α. Regarding the parameter of most interest, α, the results showed that the hierarchical Bayesian model was overall better in recovering the data-generating parameter values for high error variances than the LS-method, as indicated by the lower RMSE values in Table 1. In addition, as evident from Figure 3A the parameters estimated with the hierarchical Bayesian model were less systematically biased towards 0 or 1 than the LS-estimates, which on average, tended to overestimate the true values. In some instances, the parameters were even estimated to be at the upper boundary, independent of the true value. Although, we found very similar patterns when the α parameters of the simulated participants were drawn from a peaked Beta(15,15) distribution, contrary to what we would have expected, the accuracy of the hierarchical Bayesian model did not increase substantially. Yet, the estimates were still less biased and more accurate compared to the LS-estimates. When we inspected Figure 3B the estimates of the hierarchical Bayesian model seem to be shrunken towards the empirical mean value of .45.

Figure 3

Scatterplot of the true and estimated α parameter values



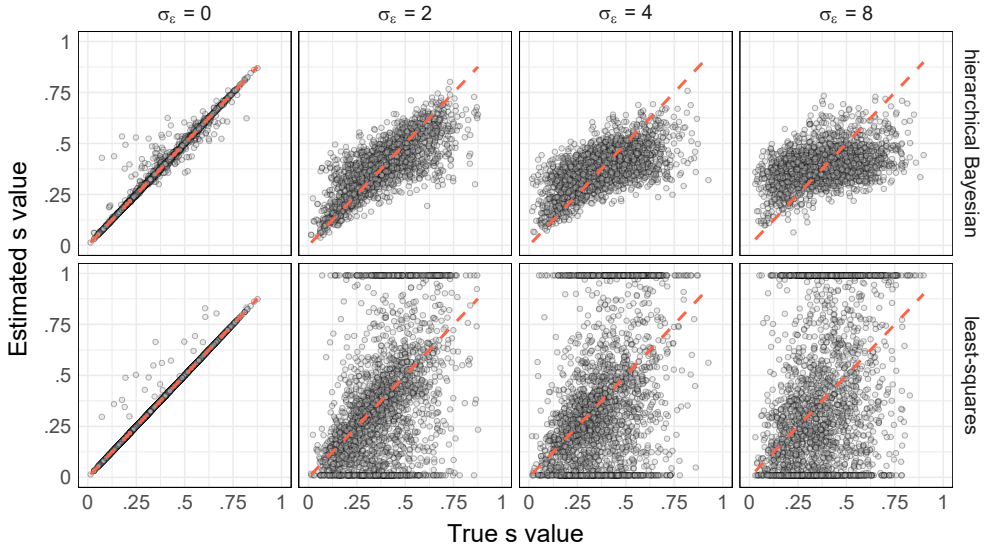
Note. The true α parameter values were drawn from a **A** Beta(1,1) or **B** Beta(15,15) distribution and for either no ($\sigma_\epsilon = 0$) or large ($\sigma_\epsilon = 8$) amounts of noise in the generated data. The two rows correspond to the different parameter estimation methods.

s. Overall, the s parameter is less well recovered than the α parameter when there is a lot

of noise in the simulated data as indicated by the higher RMSE values in Table 1. However, the hierarchical Bayesian estimates still had the lowest RMSE values. An inspection of Figure 4 suggests that the two estimation procedures show very different patterns of misestimation. The hierarchical Bayesian estimates became more clustered or shrunken (i.e., lower true values were overestimated and higher true values underestimated) towards the average of the data-generating values $M_{\text{true}} = 0.37$ when the error variance increased. The LS-estimates showed the more erratic behavior as 40.01 % of the estimates were either estimated at the lower or upper possible boundary.

Figure 4

Scatterplot of the true and estimated s parameter values for different levels of noise (σ_ϵ).



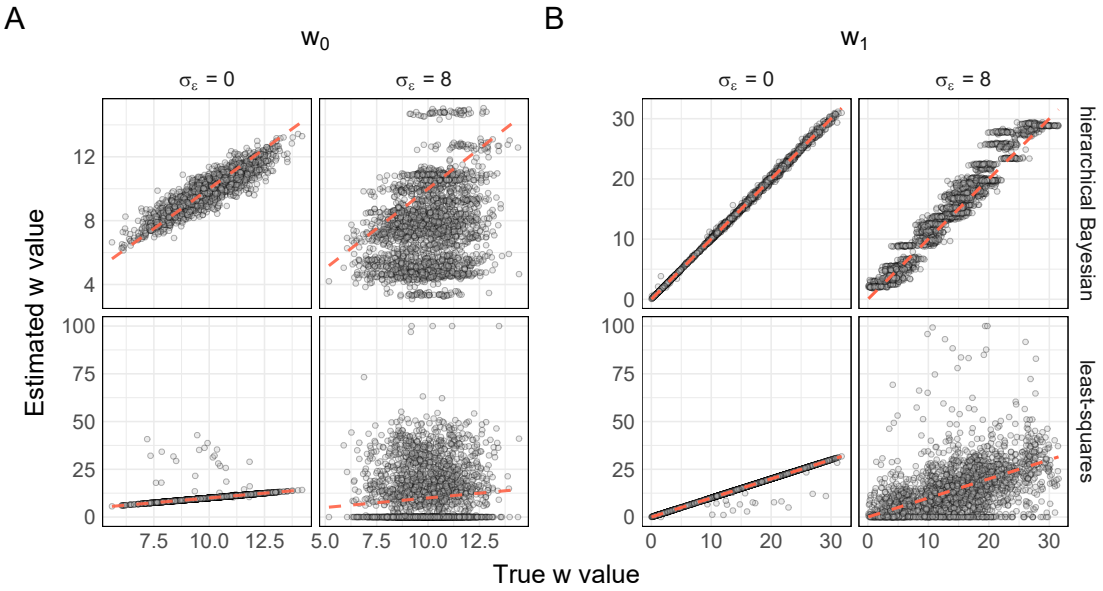
Note. The α parameter values were drawn from a Beta(1,1) distribution. The two rows correspond to the different parameter estimation methods.

wj. Both estimation methods showed a bad parameter recovery for the intercept w_0 parameter when there was a lot of noise in the simulated judgments, as indicated by the high RMSE values in Table 1 and Figure 5A. Fortunately, the cue weight parameters w_1 to w_5 (represented via w_1 in Table 1 and Figure 5B) were better recovered by both methods, with the lowest RMSE again for the hierarchical Bayesian model. Similar to the recovery of the s parameter, the estimation procedures showed very different patterns of misestimation, as evident in Figures 5B: The LS-estimates showed the tendency to estimate the parameters at the lowest possible value regardless

of the true value. In the hierarchical Bayesian model the parameter values of all 50 synthetic participants in one iteration of the simulation were estimated to have the same, or a very similar value to the other participants in the given iteration, demonstrating a strong case of shrinkage.

Figure 5

*Scatterplot of the true and estimated w_0 (**A**) and w_1 (**B**) parameter values with either no ($\sigma_\epsilon = 0$) or large ($\sigma_\epsilon = 8$) amounts of noise in the generated data.*



Note. The α parameter values were drawn from a Beta(1,1) distribution. The two rows correspond to the different parameter estimation methods.

Summary and Discussion

Overall, the results of the simulations show that the hierarchical Bayesian RulEx-J model is able to recover the underlying parameters and, as expected, doing so more accurately than the LS-approach, when there is noise in the data. However, the value of parameter recovery simulations in general can be rather limited (Lee, 2018; Lee et al., 2019). Even a model with perfect parameter recovery does not tell us that we will draw correct inferences from empirical data or that this model reflects the underlying data-generating process. Therefore, the results of this simulation serve foremost as a sanity check that the Bayesian model is correctly implemented and that the hierarchical Bayesian approach indeed leads to more accurate recovered parameters than the LS

Table 1

Root-mean-squared-error between the true and estimated parameters over all repetitions for high error variances ($\sigma_\epsilon = 8$)

Beta	Type	α	s	w_0	w_1
$a = 1, b = 1$	hB	0.10	0.15	3.48	1.95
	LS	0.28	0.36	14.53	11.00
$a = 5, b = 5$	hB	0.10	0.15	4.76	2.50
	LS	0.25	0.34	12.09	8.60
$a = 15, b = 15$	hB	0.09	0.16	4.51	2.10
	LS	0.24	0.35	10.98	7.63

Note. hB = hierarchical Bayesian, LS = Least-Squares, a and b are the shape parameters of the corresponding beta distribution from which the α parameter values were drawn.

approach that was originally used. The recovered parameter estimates of the hierarchical Bayesian approach were also less systematically biased, this is, there was not a strong tendency to over- or underestimate the true parameter values.

However, we still gain important additional insights from the simulations. The results show how the model parameters, depending on the parameter-estimation method, behave under more realistic conditions (i.e., when there is noise in the data) and what inferences we might be able to draw based on the data available. This is how informative our data are in this simulated experimental design. The observed pattern of misestimation and behavior of the hierarchical Bayesian model was more reasonable when there was a lot of noise in the data. Whereas the LS-estimates showed strong systematic biases or unpredictable erratic behavior (e.g., by estimating parameters to be on one or both of the parameter boundaries independent of the true value[11]), the

[11] The extreme cases of misestimations (i.e., parameter being estimated to be 1, regardless of the true value) for the α parameter disappeared when we relaxed the equality constraint of the s parameters of the exemplar

patterns of the hierarchical Bayesian model are demonstrations of the before mentioned shrinking property of hierarchical models (shown in Figures 3, 4 and 5). This is, the estimates are shrunken towards their corresponding group means, which in turn can lead to lower RMSE than non-hierarchical estimates (Rouder et al., 2018). This behavior is in line with previous studies that found similar results (e.g., Farrell & Ludwig, 2008). There was more shrinkage, when the synthetic participants were more similar to each other (see Figures 3A and 3B) or when there was more noise in the data (see Figures 3, 4 and 5). If there is a lot of noise in the data, these results indicate that for an experimental design with 32 trials as in the simulation, it might not be possible to achieve accurate estimates of parameter values of individuals. Given that the experimental design, the stimulus structure, and the number of trials is typical for multiple-cue judgment research, the results suggest that researchers should focus on making inferences about group-level parameters when using the hierarchical Bayesian RulEx-J Model. In order to get more precise estimates on an individual level, one has to collect more trials per participant. Figure 6 shows the difference in individual parameter-estimation accuracy for the s parameter (for $\alpha \sim \text{Beta}(15,15)$ and $\sigma_\epsilon = 8$), however, this time with 128 instead of 32 trials per participant. Increasing the number of trials increased the average correlation in simulated experiments (i.e., repetitions of the simulation) from $r = .49$ to $r = .76$.

It should also be noted that, although we report here the results for all individual level

---

module, this is, we allowed the s_i parameter of each cue i to vary freely and not be constrained to have the same value. Thus, the tendency of the rule-based model to overfit (when using LS) is due to the choice of constraining the s parameters to have the same value. Although, the recovery of the LS-estimated α parameters under high levels of noise improves when the exemplar model with free s parameters is used, the general pattern of results reported here stayed the same (i.e., hierarchical Bayesian model recovers the true parameter values more accurately under high levels of noise). The results can be found in the supplementary materials. Instead of loosening up the equality constraints on the s parameters, estimating parameters using a cross-validation approach could also prove useful, if researchers still want to use LS or ML estimations. However, as mentioned before, many studies find that exemplar models with free s parameters or attention weights show to be overly flexible and prone to overfit when using generalization tests (Hoffmann et al., 2013, 2014, 2016; von Helversen & Rieskamp, 2008, 2009)

parameters (α, s, w_j), the α parameter is the parameter of central interest and major relevance for the questions in this line of research. The results of our simulations demonstrated clearly that the hierarchical Bayesian RulEx-J model gives more precise and less biased individual estimates for the α parameter and, thus, should be preferred to alternative estimation methods.

Figure 6

Scatterplot of the true and estimated s parameter values of 30 participants with 128 trials each, for $\sigma_\epsilon = 8$ and $\alpha \sim Beta(15,5)$.

Application

In this section, we applied the hierarchical Bayesian RulEx-J model to data from three different experiments to test the validity of the α parameter, as well as to investigate if the improved model confirms previous results. First, we ran a preregistered experiment where we induced either rule-based or exemplar-based judgments from participants to validate the α parameter. Second, we reanalysed data from one of the experiments with which the original RulEx-J model was tested (Experiment 1B in Bröder et al., 2017). Third, we also reanalysed data from a different lab were the experiment showed clear differences between groups in the dominant type of judgment process used to complete the task (Experiment 1 in Trippas & Pachur, 2019). This approach allows us to show how the model can be applied to different experiments, using different stimuli, manipulations and judgment criteria. Furthermore, we can test if we are able to reproduce previous results when using the hierarchical Bayesian approach by reanalyzing data from two existing experiments, as well as testing the validity of the α parameter in a new experiment. In addition, we are able to get an idea of what effect sizes are to be expected under different

interventions manipulating the dominant mode of processing.

Data Analysis

Comparing α between conditions

All three data sets were analysed in the same way. Instead of fitting the model separately to each condition in the following experiments and then comparing the posterior means of the individual α parameters with a subsequent independent two-sample t-test, the Bayesian hierarchical approach also allows us to model these group differences directly with a slight reparameterization of the model as shown in Figure 7. This parameterization in terms of difference between group-level parameters has several advantages. First, the explicit modeling of the difference between both conditions allows us to directly implement potential theoretical assumptions and hypotheses about this difference via the prior distribution (Lee & Wagenmakers, 2013) and add potential predictors for the group difference (e.g., Bott et al., 2020; Schubert et al., 2019). Second and more importantly, Boehm et al. (2018) showed that the two-step approach of running t-tests on individual posterior estimates, can lead to incorrect conclusions and is biased towards the alternative hypothesis. To implement the parameterization in terms of group differences for the α parameter we used the following reparameterization: exp(0.5)

$$\alpha_i = \Phi(\alpha_{\text{real}_i}) \tag{7}$$

$$\alpha_{\text{real}_i} \sim \text{Normal}(\mu_{\alpha j}, \tau_\alpha) \tag{8}$$

$$\mu_{\alpha,k=1} = \mu_0 + \frac{1}{2}(\delta \times \sigma_\alpha) \tag{9}$$

$$\mu_{\alpha,k=2} = \mu_0 - \frac{1}{2}(\delta \times \sigma_\alpha) \tag{10}$$

$$\mu_0 \sim \text{Normal}(0, 1) \tag{11}$$

$$\delta \sim \text{Normal}(0, 1) \tag{12}$$

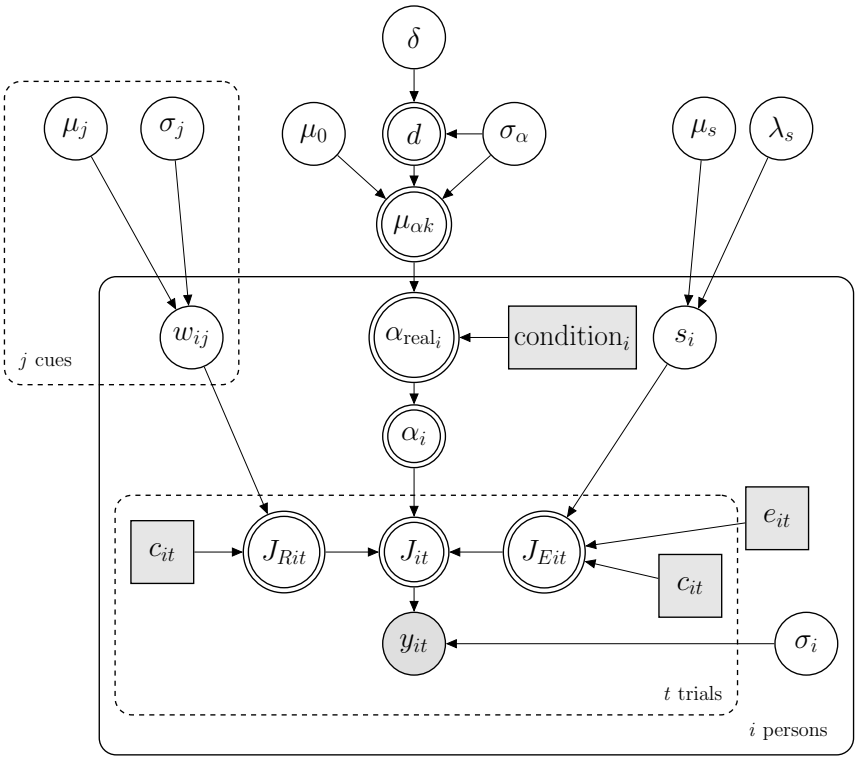
$$\tau_\alpha = \frac{1}{\sigma_\alpha^2} \tag{13}$$

$$\sigma_\alpha \sim \text{Exponential}(0.5) \tag{14}$$

The parameter μ_α reflects the overall α mean on the real scale. The parameter δ reflects the differences between both conditions on a standardized scale and hence, it reflects the effect size of the fixed effect between experimental conditions. The α value of each person i on the real scale ranging from $-\infty$ to ∞ (α_{real_i}) is then drawn from a normal distribution with a mean depending on the condition of the person with $\mu_{\alpha,j=1}$ for the rule condition and $\mu_{\alpha,j=2}$ for the exemplar condition. To get α, the α_{real_i} is then probit transformed to make sure the values are on the scale from 0 to 1.

Figure 7

Graphical model representation of the hierarchical Bayesian RulEx-J model with two-sample between-subject comparison of α.



Using this model version, we can then compute Bayes Factors based on the Savage-Dickey density ratio (SDDR, Vandekerckhove et al., 2015; Wagenmakers et al., 2010) to test hypotheses about the α parameters between conditions by computing the ratio of the prior density $p(\delta = 0|\mathcal{H}_1)$ and posterior density $p(\delta = 0|D, \mathcal{H}_1)$ at point $\delta = 0$[12]. Since we expected to find on average larger α values in the rule condition than in the exemplar condition (i.e., $\delta > 0$), we used only those

[12] The density of the posterior distribution was computed with the *dlogspline* function in the *polspline*

MCMC samples to calculate the densities that obeyed this order-restriction (Wagenmakers et al., 2010). The resulting Bayes factor of this ratio $BF_{10} = \frac{p(\delta=0|\mathcal{H}_1)}{p(\delta=0|D,\mathcal{H}_1)}$ indicates the relative evidence for $\mathcal{H}_1$ (i.e., $\delta > 0$) compared to $\mathcal{H}_0$ (i.e., $\delta = 0$, Kass & Raftery, 1995; Morey et al., 2016; Vandekerckhove et al., 2015).

For all data sets, we collected 3,000 samples from each of 3 independent MCMC chains, after 30,000 burn-in samples were discarded, 30,000 adaptive iterations, and thinning by recording every 30th sample. The convergence of the chains was checked by visual inspection and the standard $\hat{R}$ statistic ($\hat{R} < 1.02$, Gelman & Rubin, 1992). The R scripts, the JAGS models, a summary of the posterior estimates of the hyperparameters, MCMC traces, and the results files can be found in the online materials of this project.

In contrast to the parameter-recovery simulations, we used more informative prior distributions for the hyperparameters of the cue-weights μ_{w_j} to improve the convergence of the MCMC-chains. Instead of using uniform distributions, the prior distributions were centered around the cue-weight values used to generate the criterion values of the stimuli in the experiments. This is, we used prior distributions of Normal(x_j,σ) for the hyperparameters μ_{w_j}, where x_j is the cue-weight value used to generate the criterion values in the corresponding experiments (e.g., x = {10,25,20,15,13} in, Bröder et al., 2017; or x = {0.1,0.4,0.3,0.2,0.1} in Trippas & Pachur, 2019). In addition, we implemented a so-called parameter expansion for the individual cue weight parameters w_{ji} to improve the convergence of the chains (Gelman, 2006; Lee & Wagenmakers, 2013, p. 164-167) when analyzing the Bröder et al. (2017) data set, since the initial convergence of the chains was not satisfactory for these parameters in this data set. Given the different scale of criterion values in Trippas and Pachur (2019) (0-1 instead of 1-100), we also adjusted the priors for the different variance parameters (i.e., σ_i, σ_w, and Normal(μ_{w_j},σ)). The remaining prior distributions remained the same as in the parameter-recovery simulation.

---

package in R (Kooperberg, 2020)

Model comparison

In order to evaluate whether the assumption of two rather than just one of the cognitive modules is necessary, we also computed Bayes Factors per person comparing the RulEx-J model to each of the two sub-modules, this is, only rule- or exemplar-based processing. Because the two sub-modules are nested in the RulEx-J model when $\alpha = 1$ (only rule-based processing) or $\alpha = 0$ (only exemplar-based processing), we calculated the SDDR-Bayes-Factors based on the posterior distribution of α of each person.

Validation Experiment

We initially planned and ran an experiment based on the method and procedure of Bröder et al. (2017) Exp. 1A, where participants were instructed to use either a rule-based or exemplar-based strategy to solve the task. However, the manipulation did not work as expected, regardless of the analysis method used. We expect this was because we had to conduct the experiment online via Prolific due to the COVID-19 pandemic. Given the rather difficult and effortful nature of the task, we suspect that our chosen manipulation was too weak for an online setting[13]. The data can be found in the online materials of this project.

Therefore, we decided to run an additional experiment fitted to the online setting by having a simpler procedure without an extensive learning phase and a stronger manipulation. Since the main goal of this experiment was to validate and test the ability of the hierarchical Bayesian model to detect differences in the α parameter between groups or conditions, we designed an experiment where the information participants got to solve the task presumably fostered either rule- or exemplar-like processing. In the exemplar condition, we gave participants information about some exemplars, their features, and their criterion values, and instructed them that stimuli can be judged based on the similarity (i.e., the shared features) with these exemplars. In the rule condition, we informed participants that the criterion value was a linear combination of the features of the stimuli and also gave them a range of values for the criterion increases associated with each cue value.

[13] We are also not aware of other multiple-cue judgment studies which were conducted online and not in the lab.

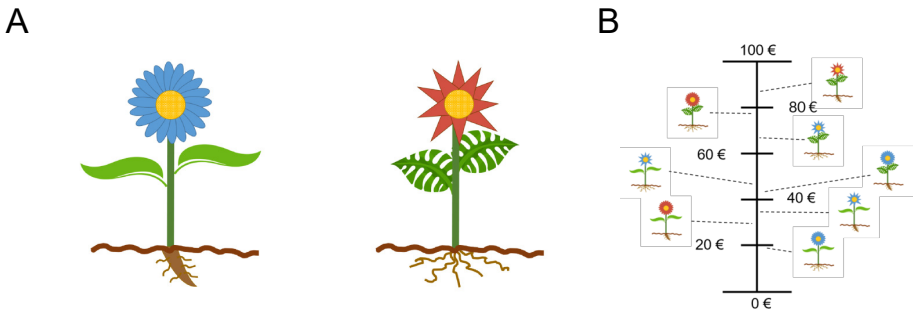
Thus, instead of instructing participants on what to learn during a learning phase as in Bröder et al. (2017) Exp. 1A (i.e., the criterion values, the cues, or a rule connecting both), we directly gave participants the information they should have learned to respond with a exemplar-based or rule-based strategy.

Method

Design and Procedure. The experiment was conducted in accordance with the ethical standards of the American Psychological Association (APA). The experiment was run online using lab.js (Henninger et al., 2021). Participants first gave their consent and then continued to read the instructions of the task. Participants were randomly assigned to one of two conditions: The exemplar ($n = 126$) or the rule condition ($n = 112$). In both conditions, the participants had to judge all 16 flowers twice, for a total of 32 trials. Depending on the condition, participants got different aids and instructions to be able to solve the task. In the exemplar condition, a visual scale (cf. Figure 8B) was presented together with the to be judged flower in each trial. The visual scale aimed to make participants base their judgments of a stimulus on the similarity with the exemplars and thus induce exemplar-based processing. For this reason, the visual scale depicted the approximate location of eight flowers (the exemplars) on a scale of prices from 0 to 100€, indicating the price of the flowers according to their cues. The participants were then told that they could judge the price of flowers according to the features and prices of the exemplary flowers depicted on the scale. For example, the left flower in Figure 8A is almost identical to the exemplar flower with the lowest price on the visual scale in Figure 8B. The only difference is the type of root (shallow or thick). In the rule condition, participants were told that the price of the flowers increased depending on the features. For instance, red flowers were more expensive than blue flowers, but the exact price increases were not known. For each of the four cues and the intercept (i.e., the price for the cheapest flower) participant received a range of possibles price increases. For instance, participant were told that red flowers cost 20 to 30€ more than blue flowers. The price ranges displayed on each trial for each of the four features and the intercept were 30 to 40€ (cue_1), 20 to 30€ (cue_2), 10 to 20€ (cue_3), 5 to 15€ (cue_4), and 7 to 13€ (intercept), respectively.

Figure 8

Example of stimuli and visual scale used in the validation experiment



Note. **A** Example of stimuli used in the validation experiment Flowers could vary on four binary cues: leave form, blossom color, petal form, and root form. **B** The visual scale shown to participants in the exemplar condition. It shows the approximate location of eight flowers (the exemplars) on a visual scale from 0 to 100€, indicating the price of the flowers according to their cues.

Hypothesis. If the manipulation of processing was successful and the α parameter of the RulEx-J model adequately reflects the process mixture, we would expect substantially higher α parameter estimates in the rule condition than in the exemplar condition. Hence, we expected to find a $\delta > 0$ which indicates a higher average α level of the rule condition compared to the exemplar condition.

Materials and Measures. Participants were presented with 16 flowers and asked to judge the price of each flower on a scale from 0 to 100. Each flower was characterized by four binary cues, which corresponded to four features (cue_1 : leaf form, cue_2 : blossom color, cue_3 : petal form, cue_4 : root form). Two examples are shown in Figure 8A. The criterion values were computed via a linear function of the form $\text{Criterion} = 10 + 32\text{cue}_1 + 27\text{cue}_2 + 18\text{cue}_3 + 9\text{cue}_4$. The assignment of cues and cue values to the features was the same for each participant.

Participants. In total we collected data from $N = 266$ participants who completed the study via university mailing lists ($n = 45$) and Prolific Academic ($n = 221$) [14]. As preregistered, we

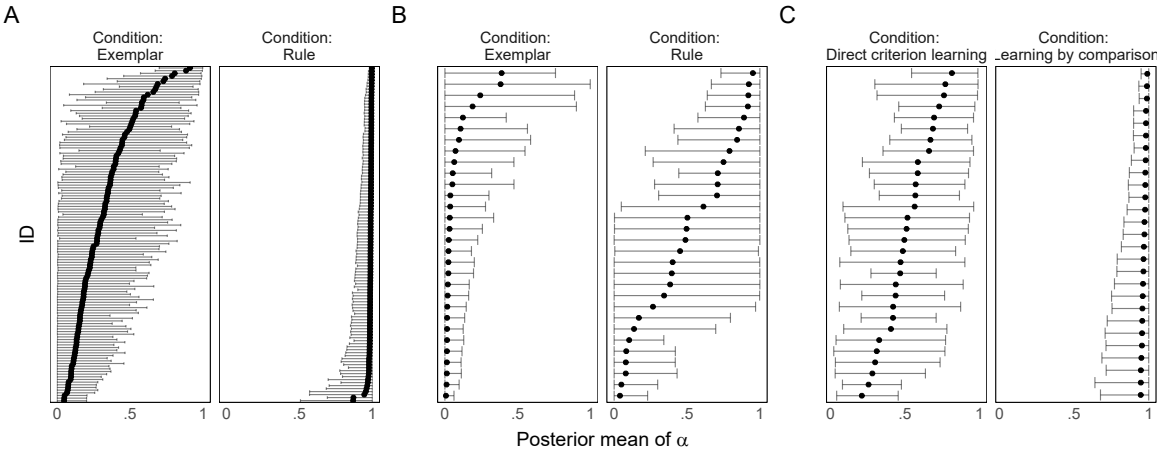
[14] We initially wanted to collect participants only via university mailing lists, however, due to very slow recruitment because to the COVID Pandemic we decided to also recruit participants via Prolific Academic Ltd.

excluded $n = 4$ participants who indicated that their data should not be used for data analysis (Aust et al., 2013). Furthermore, since it was important that participants understood all instructions clearly, we also decided to exclude $n = 5$ participants who indicated that they did not speak German fluently. In a last step, we excluded $n = 19$ participants who had an RMSE greater than 25 between their judgments and the actual criterion values, which indicated that they did not follow the instructions[15]. Our final sample thus consisted of $N = 238$ participants (117 female, 4 non-binary, mean age = 29.87, $SD = 9.88$).

Results

Figure 9

The posterior means of α with the corresponding 95% credibility intervals (CI) for each participant in both conditions.



Note. **A** the new validation experiment, **B** Experiment 1B of Bröder et al. (2017), **C** Experiment 1B of Trippas & Pachur (2019).

Difference in α between conditions. The posterior distribution of δ, as shown in Figure 10, had a mean of 3.62 ($SD = 0.40$, 95%-CI = [2.90,4.47]). The Bayes-Factor indicates that the hypothesis of having larger α values in the rule condition (or $\delta > 0$, $\mathcal{H}_1$) is $\text{BF}_{10} > 1000$ times

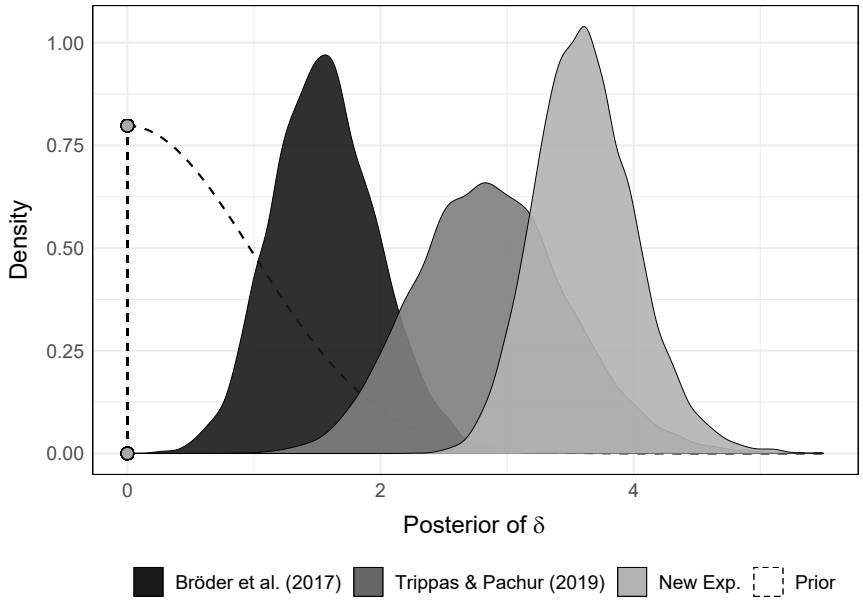
[15] We did not preregister the last two filtering steps (i.e., based on language and RMSE). However, the results presented in this section do not change substantially, when the excluded participants were included.

more likely than the hypothesis that there is no difference in α between the conditions ($\mathcal{H}_0$)[16]. The posterior means of the individual α's with the corresponding 95%-credibility-intervals (CI) for each participant in both conditions are shown in Figure 9A.

Model comparison. The results of model comparison analysis on an individual level are shown in Table 2. Most participants in the exemplar condition were best described by the RulEx-J model (58.73%) and then by the exemplar model (37.30%). In the rule condition, the rule model fitted best for most participants (53.57%) compared to the RulEx-J model (45.54%).

Figure 10

Prior and posterior distribution of the effect size δ for the hierarchical Bayesian analysis.



Note. The markers highlight the densities at $\delta = 0$ used to estimate the Bayes factor.

Bröder et al. (2017)

Given the rather technical nature of the validation experiment without the typical learning phase and a direct manipulation of the α parameter, we also applied the hierarchical Bayesian RulEx-J model to a more realistic data set, which was used in the original RulEx-J paper by

[16] The results of the analysis using least-squares estimation can be found in the online supplementary material

Table 2

Proportion of best fitting model for each person as determined by the SDDR-Bayes-Factor

Experiment	Condition	% RulEx-J	% Rule	% Exemplar
Validation Exp.	exemplar	58.73	3.97	37.30
	rule	45.54	53.57	0.89
Bröder et al. (2017)	exemplar	16.67	10.00	73.33
	rule	50.00	23.33	26.67
Trippas & Pachur (2019)	dcl	70.00	26.67	3.33
	lbc	40.74	55.56	3.70

Note. dcl = direct criterion learning, lbc = learning by comparison.

Bröder et al. (2017). In this experiment, the 60 participants had to judge the severity of a patient's disease on a scale from 0 to 100, based on a set of four binary symptoms (e.g., fever vs. hypothermia). The experiment itself consisted of four phases, a memorization phase, a learning phase, a decision phase, and a final testing phase. However, the decision phase and its data are not important for this reanalysis, since the focus of our work lies on the judgment data. Since the experiment focused on memory-based judgments, in the memorization phase participants had to learn the cues from 14 of 16 patients (the two most extreme patterns were left out) until they remembered 80% of the cues correctly. In the training phase, participants then had to judge the severity of illness of eight patients (the exemplars). They then received feedback about the actual criterion value after their judgment. For the experimental manipulation, participants were instructed to either use the feedback about the correct criterion values to learn a mathematical rule connecting cue and criterion values (rule condition) or to memorize the patients and their respective criterion values (exemplar condition). The training phase consisted of eight blocks with eight trials each (one for each exemplar). In the final testing phase, the participants had to judge the criterion values of all 16 patients. Depending on the condition, they were instructed to either apply the mathematical rule they learned earlier (rule condition) or judge untrained objects by

their similarity to the memorized objects (exemplar condition). The results in the original study were based on least-squares estimation and showed that the average α parameter was larger in the rule condition ($M = .60$, $SD = .30$) than in the exemplar condition ($M = .39$, $SD = .23$). By reanalyzing the data with the Bayesian hierarchical RulEx-J model, we expected to replicate this result, this is, $\delta > 0$ when directly modeling group differences in α.

Results

Difference in α between conditions. The δ parameter of the group-difference RulEx-J model had a posterior mean of 1.57 ($SD = 0.42$, 95%-CI = [0.80,2.44]). The Bayes factor of $BF_{10} = 367.15$ indicated extreme evidence for the alternative hypothesis which assumed a difference in the α parameter between conditions (i.e., $\delta > 0$). Again, Figure 9B shows the posterior means of the estimated α parameters with the corresponding 95%-CI for each participant in both conditions.

Model comparison. For most participant in the rule condition the RulEx-J model was the best fitting model (50.00%), but in the exemplar condition the exemplar model was better describing the behavior of more participants (73.33%) than the RulEx-J model (16.67%, see Table 2).

Trippas & Pachur (2019)

To supplement our analyses with data from another lab, we reanalysed data from Experiment 1B from Trippas and Pachur (2019). In a series of well-designed experiments Trippas and Pachur (2019) investigated why people's reliance on rule-based and exemplar-based processing as well as generalization ability differs substantially between two types of learning tasks: direct criterion learning (dcl) and learning by comparison (lbc). In their experiments Trippas and Pachur (2019) used 15 toxic bugs as stimuli, which could differ in four binary cues and vary in their toxicity level between 0 and 1. In Experiment 1 participants were randomly assigned to one of three conditions: learning by comparison, direct criterion learning, or direct criterion learning with a reference object. However, for our purpose we only focus on the first two conditions (dcl and lbc), which led to the greatest differences in what strategy was used. Each condition consisted of $n = 30$

participants. In the training phase of the direct criterion learning condition, in each trial participants had to judge if a presented bug was deadly (i.e., had a toxicity level higher than .5) or not. After each decision, participants got feedback indicating if their decision was correct or not, as well as the exact toxicity level of the bug. In the learning by comparison condition, participants were presented with two bugs in each trial and asked to decide which was more toxic. After each trial, participants got again feedback about the correctness of their response, but not about the exact toxicity level. In both of the conditions, the same 10 out of the 15 possible bugs were used as exemplars. After the training phase, participants in both conditions had to estimate the continuous toxicity level of each of the 15 bugs in the testing phase. For more detailed information about the experiment see Trippas and Pachur (2019). Among other things, strategy classification via model comparison showed that 27 out of 30 participants (90 %) in the learning by comparison condition but only 10 out of 30 participants (33 %) in the direct criterion learning condition were best described by a rule-based strategy. When reanalyzing the data with the Bayesian hierarchical RulEx-J model, we therefore expect to find higher α values in the learning by comparison condition compared to the direct criterion learning condition, this is, $\delta > 0$ when modeling group differences in α directly.

Results

Difference in α between conditions. The posterior distribution of the standardized effect parameter δ of the group-difference RulEx-J model had a mean of 2.88 (SD = 0.60, 95%-CI = [1.77,4.10], see Figure 10B). The SDDR-Bayes-factor of $BF_{10} > 1000$ indicated extreme evidence for the hypothesis that the average α parameter is higher in the learning by comparison condition (i.e., $\delta > 0$) compared to the hypothesis of having no difference (i.e., $\delta = 0$). Estimates of the individual α parameters [17] are shown in Figure 9C.

Model comparison. The judgments of most participants in the dcl condition were best described by the RulEx-J model (70.00%). However, in the lbc condition the rule-only model described the responses of more participants better (55.56%) than the RulEx-J model (40.74%).

[17] When fitting the RulEx-J model, we excluded three participants from the lbc condition, since their perfect performance in the judgment task made the model not converge for these participants.

Discussion and Summary

We presented results of a new experiment, demonstrating the validity of the α parameter of the RulEx-J model to measure differences in rule-based and exemplar-based processing between conditions. We further showed that with the hierarchical Bayesian RulEx-J model we were able to reproduce the results of previous experiments of different research when comparing the α between conditions. Hence, the experiments demonstrate that modeling the data with the improved RulEx-J implementation yields meaningful results in terms of the parameters estimating the mixture of the processes.

The results of the individual model comparisons showed that overall experiments the RulEx-J model best described the judgments of most participants (46.95 %) compared to the two simpler sub-process, this is pure exemplar- (24.20 %) or pure rule-based processing (28.85 %). However, these results also show that there are some individual differences. The responses of a substantial number of participants were better described by the simpler sub-process model of the corresponding conditions (i.e., the rule model in the rule/lbc condition, or the exemplar model in the exemplar/dcl condition), or sometimes even the other way around. Thus it seems that the additional complexity of the RulEx-J model does not always pay-off in terms of model fit and probably depends on how easy it is to learn and apply the underlying rule (e.g., in the validation experiments) or how well participants are able to learn all exemplars and the corresponding criterion values (e.g., in Bröder et al. (2017) there was an additional memorization phase to learn all exemplars). Since the RulEx-J model is foremost intended as a measurement model, which includes the possibility of pure rule- or exemplar-based processing and the α values between the conditions in the analysed experiments reflect the expected differences in processing mode, this is not a problem for the RulEx-J model.

In addition, in the simulations in the previous section, we tested the ability of the hierarchical Bayesian RulEx-J model to recover parameter values under different levels of noise. The application of the model to these different data sets allows us to get estimates about levels of noise that could be expected in real data. According to the model implementation we used here, the responses of participants in a given trial are modeled as $y \sim \text{Normal}(J_{it}, \sigma_i)$. Using the

posterior mean of σ_i of each person as a (model-based) estimate of the noise in the data, we found a median noise level of $\hat{\sigma_\epsilon} = 8.64$, ranging from 0.9 to 37. From all 355 participants in all experiments, 2.82% had $\sigma_i < 2$, 9.86% had $\sigma_i < 4$, and 41.41% had $\sigma_i < 8$. Therefore, our chosen levels of noise in the simulation were not unrealistic, although a bit too optimistic. However, these results show that the median empirically observed levels of noise over three experiment with typical stimuli and typical trial sizes, are actually similar to the highest levels of noise considered in the simulation. The simulation results showed that for these apparently realistic levels of noise there were clear deficits in the recovery of the underlying parameters of individual participants when using the traditional LS approach. Thus, researchers should refrain from making inferences based on individual-level parameter estimates under these circumstances. The hierarchical Bayesian model fares better than the LS approach, but, based on the simulation, estimated parameters of individual participants should still be used with care when noise levels are high.

General Discussion

In this article, we introduced a hierarchical Bayesian implementation of the RulEx-J model. Simulation studies showed that the hierarchical Bayesian RulEx-J model is able to recover parameters more accurately and less biased than a separate analysis of individuals with a least-squares estimation. This advantage of the hierarchical Bayesian implementation became especially clear when there was noise in the data. The individual α parameters, which measure the relative impact of rule- and exemplar-based processes on the final judgment and thus are the parameters of most interest, were recovered reasonably well, even when there was substantial noise in the data. Due to the hierarchical structure, individual s and cue-weight parameters w_j were recovered less accurately with increasing noise and, thus, increasing shrinkage. However, group-level inferences are still possible. These findings are in line with other simulation studies comparing hierarchical and non-hierarchical Bayesian and maximum-likelihood based estimation methods (e.g., Farrell & Ludwig, 2008). Furthermore, a new experiment where the information participants got to solve the task lead to a rule- or exemplar-like processing added evidence to the validity of the α parameter, as well as to the validity of the Bayesian hierarchical RulEx-J model. In addition, we showed that we could reproduce the results of two previous studies with the hierarchical Bayesian

implementation of the RulEx-J model by directly incorporating group-differences in our model. As already suggested by Boehm et al. (2018), this approach is more viable than a two-step analysis approach (i.e., estimating individual parameters and then computing a subsequent t-test), since the different variance in the individual α parameter estimates may be due to different levels of shrinkage, which in turn would bias inferences.

Limitations and future directions

In our second simulation we induced noise to the judgments by adding normally distributed error to the generated judgments. While this mimics general noise present in real experimental data due to various influences, there are other error or contamination processes present in real experiments, which might influence the ability of the model to recovery parameters in unique ways, such as guessing, biased responding, or the use of other judgment strategies. Second, from the simulation results it seems that the model needs a large number of individual data points to get precise individual estimates (especially for the s and w_j) the more noise there is in the data. However, in practice the number of individual data points research could get might often be limited by the typical multiple-cue judgment paradigm itself, where individual participants have to learn the cues and criterion values, as well as their relationship, of several stimuli. Dependent on what the participants have to learn, it might not be possible to increase the number of cues or stimuli without having losses in performance. Third, we did not run extensive prior sensitivity analysis for each analysis. However, since the results did not change in the cases where we tried different prior specifications, we are confident that our results are robust for different reasonable prior distributions.

While the state-of-the-art Bayesian hierarchical approach improves upon problems of parameter estimation of the original RulEx-J model as a measurement model, the Bayesian framework used in this article also offers new possibilities to implement and then compare different model variants to answer theoretical questions. For instance, by incorporating a learning process (e.g., Hoffmann et al., 2019), adding possible contamination processes (e.g., Zeigenfuse & Lee, 2010), more complex rule- or exemplar-process models (e.g., Izydorczyk & Bröder, 2021), integrating additional sources of information or covariates (e.g., mouse-tracking, eye-tracking, EEG).

Currently, the RulEx-J model is foremost intended as a pragmatic measurement tool and thus might not describe the actual cognitive processes that lead to a judgment. Although the empirical evidence presented above makes it plausible that there is indeed a mixture between rule- and exemplar-based process involved when people make their judgments, there are possible other conceptualizations how rule-based and exemplar-based processes interact. A remaining challenge to establish the RulEx-J model as a more epistemic cognitive model is to test and compare different theoretical conceptualizations of the process mixing. Instead of having a constant mixture of both processes at all times, it might be possible that participants vary the relative proportion of processes between trials, or switch between processes over sequences of trials (Lee & Gluck, 2020; Lee et al., 2019), trial-by-trial, or even between stimuli (as assumed by the ATRIUM model, Erickson & Kruschke, 1998). Other mixture processes might also be possible, such as the one proposed by the CX-COM (combining Cue abstraction with eXemplar memory assuming COMpetitive memory retrieval, Albrecht et al., 2019) model. The CX-COm model proposes a two-step process were one exemplar is recalled competitively from a set of exemplars and its associated criterion value (i.e., the initial judgment) is then adjusted based on abstracted cue knowledge. We are convinced that the improved modeling approach presented here offers a start to address these hitherto unanswered research questions.

References

Albrecht, R., Hoffmann, J., Pleskac, T., 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*. https://doi.org/10.1037/xlm0000772

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, *120*(1), 3–19. https://doi.org/10.1037/0096-3445.120.1.3

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown* [R package version 0.1.0.9942]. https://github.com/crsh/papaja

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*(2), 527–535. https://doi.org/10.3758/s13428-012-0265-2

Boehm, U., Marsman, M., Matzke, D., & Wagenmakers, E.-J. (2018). On the importance of avoiding shortcuts in applying cognitive models to hierarchical data. *Behavior Research Methods*, *50*(4), 1614–1631. https://doi.org/10.3758/s13428-018-1054-3

Bott, F. M., Heck, D. W., & Meiser, T. (2020). Parameter validation in hierarchical MPT models by functional dissociation with continuous covariates: An application to contingency inference. *Journal of Mathematical Psychology*, 14.

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*(4), 406–417. https://doi.org/10.1080/20445911.2018.1444613

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.

Bröder, A., Newell, B. R., & Platzer, C. (2010). Cue integration vs. exemplar-based reasoning in multi-attribute decisions from memory: A matter of cue representation. *Judgment and Decision Making*, *5*(5), 326–338.

Brooks, L. R., & Hannah, S. D. (2006). Instantiated features and the use of "rules." *Journal of Experimental Psychology: General*, *135*(2), 133.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*, 193–217.

Corporation, M., & Weston, S. (2019). *Dosnow: Foreach parallel adaptor for the 'snow' package* [R package version 1.0.18]. https://CRAN.R-project.org/package=doSNOW

Denwood, M. J. (2016). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, *71*(9), 1–25. https://doi.org/10.18637/jss.v071.i09

Eddelbuettel, D., & Balamuta, J. J. (2017). Extending extitR with extitC++: A Brief Introduction to extitRcpp. *PeerJ Preprints*, *5*, e3188v1. https://doi.org/10.7287/peerj.preprints.3188v1

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. https://doi.org/10.18637/jss.v040.i08

Efron, B., & Morris, C. (1977). Stein's Paradox in Statistics. *Scientific American*, *236*(5), 119–127. https://doi.org/10.1038/scientificamerican0577-119

Einhorn, H. J., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Linear Regression and Process-Tracing Models of Judgment. *Psychological Review*, *86*(5), 465–485. https://doi.org/10.1037/0033-295X.86.5.465

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

Farrell, S., & Ludwig, C. J. H. (2008). Bayesian and maximum likelihood estimation of hierarchical response time models. *Psychonomic Bulletin & Review*, *15*(6), 1209–1217. https://doi.org/10.3758/PBR.15.6.1209

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815. https://doi.org/10.1080/0266476042000214501

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (3nd ed.). CRC Press/Taylor & Francis Group.

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

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

Hannah, S. D., & Brooks, L. R. (2009). Featuring familiarity: How a familiar feature instantiation influences categorization. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *63*(4), 263–275. https://doi.org/10.1037/a0017919

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2021). Lab.js: A free, open, online study builder. *Behavior Research Methods*. https://doi.org/10.3758/s13428-019-01283-5

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*(2), 233–249. https://doi.org/10.1002/bdm.1958

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, &amp Computers*, *16*(2), 96–101. https://doi.org/10.3758/BF03202365

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2013). Deliberation's blindsight: How cognitive load can improve judgments. *Psychological Science*, *24*(6), 869–879. https://doi.org/10.1177/0956797612463581

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2014). General pillars of judgment: How memory abilities affect performance in rule-based and exemplar-based judgments. *Journal of Experimental Psychology*, *143*, 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. https://doi.org/10.1037/xlm0000241

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2019). Testing learning mechanisms of rule-based judgment. *Decision*, *6*(14), 305–334. https://doi.org/https://doi.org/10.1037/dec0000109

Izydorczyk, D., & Bröder, A. (2021). Exemplar-based judgment or direct recall: On a problematic procedure for estimating parameters in exemplar models of quantitative judgment. *Psychonomic Bulletin & Review*, 1–19.

Izydorczyk, D., & Bröder, A. (2022, July 28). Measuring%20the%20mixture%20of%20rule-based%20and%20exemplar-based%20processes%20in%20judgment:%20A%20hierarchical%20Bayesian%20approach.%20Retrieved%20from%20osf.io/7mabe.

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition*, *106*(1), 259–298. https://doi.org/10.1016/j.cognition.2007.02.003

Juslin, P., Olsson, H., & Olsson, A. C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology*, *132*(1), 133–156. https://doi.org/10.1037/0096-3445.132.1.133

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science*, *26*(5), 563–607. https://doi.org/10.1016/S0364-0213(02)00083-6

Karlsson, L., Juslin, P., & Olsson, H. (2008). Exemplar-based inference in multi-attribute decision making: Contingent, not automatic, strategy shifts? *Judgment and Decision Making*, *3*(3), 244–260.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

Katahira, K. (2016). How hierarchical models improve point estimates of model parameters at the individual level. *Journal of Mathematical Psychology*, *73*, 37–58. https://doi.org/10.1016/j.jmp.2016.03.007

Kooperberg, C. (2020). *Polspline: Polynomial spline routines* [R package version 1.1.19]. https://CRAN.R-project.org/package=polspline

Lee, M. D. (2018, March 23). Bayesian Methods in Cognitive Modeling. In J. T. Wixted (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (pp. 1–48). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119170174.epcn502

Lee, M. D., & Gluck, K. A. (2020). Modeling Strategy Switches in Multi-attribute Decision Making. *Computational Brain & Behavior*. https://doi.org/10.1007/s42113-020-00092-w

Lee, M. D., Gluck, K. A., & Walsh, M. M. (2019). Understanding the complexity of simple decisions: Modeling multiple behaviors and switching strategies. *Decision*, *6*(4), 335–368. https://doi.org/10.1037/dec0000105

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

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*(4), 330–354. https://doi.org/10.1006/jesp.1998.1353

Mata, R., von Helversen, B., Karlsson, L., & Cüpper, L. (2012). Adult age differences in categorization and multiple-cue judgment. *Developmental Psychology*, *48*(4), 1188–1201. https://doi.org/10.1037/a0026084

Mattes, A., Tavera, F., Ophey, A., Roheger, M., Gaschler, R., & Haider, H. (2020). Parallel and serial task processing in the PRP paradigm: A drift–diffusion model approach. *Psychological Research*. https://doi.org/10.1007/s00426-020-01337-w

McElreath, R. (2020). *McElreath, R: Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (2nd ed.). CRC Press/Taylor & Francis Group.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Mersmann, O., Trautmann, H., Steuer, D., & Bornkamp, B. (2018). *Truncnorm: Truncated normal distribution* [R package version 1.0-8]. https://CRAN.R-project.org/package=truncnorm

Microsoft & Weston, S. (2020). *Foreach: Provides foreach looping construct* [R package version 1.5.0]. https://CRAN.R-project.org/package=foreach

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18. https://doi.org/10.1016/j.jmp.2015.11.001

Müller, K., & Wickham, H. (2020). *Tibble: Simple data frames* [R package version 3.0.1]. https://CRAN.R-project.org/package=tibble

Nosofsky, R. M. (1984). Choice, similarity and the context theory of classification. *Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104–114. https://doi.org/10.1037/0278-7393.10.1.104

Pachur, T., & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology*, *65*(2), 207–240. https://doi.org/10.1016/j.cogpsych.2012.03.003

Persson, M., & Rieskamp, J. (2009). Inferences from memory: Strategy- and exemplar-based judgment models compared. *Acta Psychologica*, *130*(1), 25–37. https://doi.org/10.1016/j.actpsy.2008.09.010

Platzer, C., & Bröder, A. (2012). Most people do not ignore salient invalid cues in memory-based decisions. *Psychonomic Bulletin & Review*, *19*(4), 654–661. https://doi.org/10.3758/s13423-012-0248-4

Plummer, M. (2003). JAGS: A program for analysis of bayesian graphical models using gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing*.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Regehr, G., & Brooks, L. R. (1993). Perceptual manifestations of analytic structure: The priority of holistic individuation. *Journal of Experimental Psychology: General*, *122*, 92–144.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573–604. https://doi.org/10.3758/BF03196750

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*(2), 195–223. https://doi.org/10.3758/BF03257252

Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal Detection Models with Random Participant and Item Effects. *Psychometrika*, *72*(4), 621–642. https://doi.org/10.1007/s11336-005-1350-6

Rouder, J. N., Morey, R. D., & Pratte, M. S. (2018). Bayesian hierarchical models of cognition. In W. H. Batchelder, H. Colonius, E. N. Dzhafarov, & J. Myung (Eds.), *New Handbook of Mathematical Psychology* (pp. 504–551). Cambridge University Press. https://doi.org/10.1017/9781139245913.010

Scheibehenne, B., & Pachur, T. (2015). Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice. *Psychonomic Bulletin and Review*, *22*(2), 391–407. https://doi.org/10.3758/s13423-014-0684-4

Schlegelmilch, R., & von Helversen, B. (2020). The influence of reward magnitude on stimulus memory and stimulus generalization in categorization decisions. *Journal of Experimental Psychology: General*, *149*(10), 1823–1854. https://doi.org/10.1037/xge0000747

Schubert, A.-L., Nunez, M. D., Hagemann, D., & Vandekerckhove, J. (2019). Individual Differences in Cortical Processing Speed Predict Cognitive Abilities: A Model-Based Cognitive Neuroscience Account. *Computational Brain & Behavior*, *2*(2), 64–84. https://doi.org/10.1007/s42113-018-0021-5

Shiffrin, R., Lee, M., Kim, W., & Wagenmakers, E.-J. (2008). A Survey of Model Evaluation Approaches With a Tutorial on Hierarchical Bayesian Methods. *Cognitive Science: A Multidisciplinary Journal*, *32*(8), 1248–1284. https://doi.org/10.1080/03640210802414826

Steingroever, H., Pachur, T., Šmíra, M., & Lee, M. D. (2018). Bayesian techniques for analyzing group differences in the Iowa Gambling Task: A case study of intuitive and deliberate decision-makers. *Psychonomic Bulletin & Review*, *25*(3), 951–970. https://doi.org/10.3758/s13423-017-1331-7

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. https://doi.org/10.1037/xlm0000696

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015, December 10). *Model Comparison and the Principle of Parsimony* (J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels, Eds.; Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199957996.013.14

van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E.-J. (2011). Cognitive model decomposition of the BART: Assessment and application. *Journal of Mathematical Psychology*, *55*(1), 94–105. https://doi.org/10.1016/j.jmp.2010.08.010

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., Karlsson, L., Rasch, B., & Rieskamp, J. (2014). Neural substrates of similarity and rule-based strategies in judgment. *Frontiers in Human Neuroscience*, *8*, 1–13. https://doi.org/10.3389/fnhum.2014.00809

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*(1), 220–229. https://doi.org/10.1037/a0016690

von Helversen, B., & Rieskamp, J. (2008). The mapping model: A cognitive theory of quantitative estimation. *Journal of Experimental Psychology*, *137*(1), 73–96. https://doi.org/10.1037/0096-3445.137.1.73

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*(4), 867–889. https://doi.org/10.1037/a0015501

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *Dplyr: A grammar of data manipulation* [R package version 1.0.0]. https://CRAN.R-project.org/package=dplyr

Wirebring, L. K., Stillesjö, S., Eriksson, J., Juslin, P., & Nyberg, L. (2018). A similarity-based process for human judgment in the parietal cortex. *Frontiers in Human Neuroscience*, *12*, 1–18. https://doi.org/10.3389/fnhum.2018.00481

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd) [ISBN 978-1498716963]. Chapman; Hall/CRC. https://yihui.org/knitr/

Zeigenfuse, M. D., & Lee, M. D. (2010). A general latent assignment approach for modeling psychological contaminants. *Journal of Mathematical Psychology*, *54*(4), 352–362. https://doi.org/10.1016/j.jmp.2010.04.001

Psychonomic Bulletin and Review
https://doi.org/10.3758/s13423-020-01861-1

THEORETICAL REVIEW

Exemplar-based judgment or direct recall: On a problematic procedure for estimating parameters in exemplar models of quantitative judgment

David Izydorczyk[1] · **Arndt Bröder**[1]

Accepted: 5 December 2020
© The Author(s) 2021

Abstract

Exemplar models are often used in research on multiple-cue judgments to describe the underlying process of participants' responses. In these experiments, participants are repeatedly presented with the same exemplars (e.g., poisonous bugs) and instructed to memorize these exemplars and their corresponding criterion values (e.g., the toxicity of a bug). We propose that there are two possible outcomes when participants judge one of the already learned exemplars in some later block of the experiment. They either have memorized the exemplar and their respective criterion value and are thus able to recall the exact value, or they have not learned the exemplar and thus have to judge its criterion value, as if it was a new stimulus. We argue that psychologically, the judgments of participants in a multiple-cue judgment experiment are a mixture of these two qualitatively distinct cognitive processes: judgment and recall. However, the cognitive modeling procedure usually applied does not make any distinction between these processes and the data generated by them. We investigated potential effects of disregarding the distinction between these two processes on the parameter recovery and the model fit of one exemplar model. We present results of a simulation as well as the reanalysis of five experimental data sets showing that the current combination of experimental design and modeling procedure can bias parameter estimates, impair their validity, and negatively affect the fit and predictive performance of the model. We also present a latent-mixture extension of the original model as a possible solution to these issues.

Keywords Judgment · Exemplar model · Recall

In their everyday lives, people have to make judgments of different importance in a variety of domains and situations. For instance, customers in a restaurant have to predict how tasty a meal will be, doctors have to judge the severity of a patient's disease, and employers have to judge how well a possible employee will perform in the future. Such judgments require inferring a continuous criterion (e.g., tastiness) from a number of cues (e.g., is there cheese on it or not) of a given judgment object.

One process people may rely on to make these judgments is based on previously encountered objects and their criterion values stored in memory (Juslin, Olsson, & Olsson, 2003; Juslin, Karlsson, & Olsson, 2008). New objects are then judged based on the similarity to these exemplars (Juslin et al., 2003). For example, a diabetic person needs to judge the amount of carbohydrates in a dish to estimate the amount of insulin she has to apply. To do so when confronted with a new meal, she might think of previous meals (i.e., the memorized exemplars) and compare them to the current meal. The amount of carbohydrates of the new meal will then be judged according to the similarity of this new meal to past meals in memory and their respective amount of carbohydrates (i.e., the criterion value of the exemplars), whereby more similar past meals will have a stronger influence on the judgment than dissimilar ones.

Such a judgment strategy is usually described by exemplar models (e.g., Juslin et al., 2003; von Helversen & Rieskamp, 2008). Exemplar models have originally been

This research was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG, GRK 2277) to the Research Training Group "Statistical Modeling in Psychology" (SMiP).

✉ David Izydorczyk
izydorczyk@uni-mannheim.de

[1] Department of Psychology, Experimental Psychology Lab, School of Social Sciences, University of Mannheim, D-68131 Mannheim, Germany

Published online: 09 June 2021

used in many different domains such as memory (e.g., Hintzman, 1984) and categorization and classification (e.g., Medin & Schaffer, 1978; Nosofsky, 1984). One exemplar model originally used for modeling stimulus categorization, the Context Model (Medin & Schaffer, 1978), has been extended to account for data of continuous judgments from multiple cues (Juslin et al., 2003; Juslin & Persson, 2002). During the last two decades, this or related exemplar models have been used to describe one possible cognitive process in studies of multiple-cue judgments as an important area in judgment and decision-making (JDM) research (e.g., Bröder & Gräf, 2018; Hoffmann, von Helversen, & Rieskamp, 2013; Juslin et al., 2003; Mata, von Helversen, Karlsson, & Cüpper, 2012; Pachur & Olsson, 2012; Persson & Rieskamp, 2009; von Helversen & Rieskamp, 2009).

In the current paper, we argue that the usual practice of how these exemplar models are used in multiple-cue judgment research in combination with the paradigm commonly used in this field of research leads to biased estimation and impaired validity of parameters. We claim that the problems we tackle here are particularly pronounced in the multiple-cue judgment literature as compared to categorization research where the model and the experimental paradigm originated (Juslin et al., 2003). We highlight a severe problem of the application of these exemplar models in JDM research in the following respects: First, we will briefly describe the typical experimental paradigm and modeling procedure used in multiple-cue judgment studies and introduce the context model (Medin & Schaffer, 1978) as an example for an exemplar-based model how it is used in this line of research (e.g., Bröder & Gräf, 2018; Juslin et al., 2003; von Helversen & Rieskamp, 2008; Wirebring, Stillesjö, Eriksson, Juslin, & Nyberg, 2018). We will then illustrate how this paradigm with the currently applied specification of the exemplar model can potentially distort parameter estimation. Second, we will present simulation results demonstrating biased estimation and impaired validity of parameters. In addition, since many multiple-cue judgment studies use model-fit indices like the RMSE to compare different judgment process models (e.g., Hoffmann, von Helversen, & Rieskamp, 2014; Mata et al., 2012; Wirebring et al., 2018) we will also look at the model fit and predictive performance of the model. Third, we will present the results of five reanalyzed experimental data sets demonstrating that this problem also threatens the interpretation of real data. As a remedy, we will present a latent-mixture extension of the original model, as a possible solution to these problems. Although the focus of this paper lies on the multiple-cue judgment literature and how the exemplar models are applied there, we will also discuss if and how the results of this work might extend to other research areas where these exemplar models are applied.

Typical design and estimation procedure in multiple-cue judgment studies

A typical experiment in the multiple-cue judgment literature employing exemplar models consists of at least two phases: a training phase and a testing phase. In the training phase, participants have to learn the cues and the criterion values of some stimuli (i.e., the exemplars), as well as the relationship between the cues and the criterion. This is typically done by repeatedly judging the criterion values of a sample of objects and receiving trial-by-trial-feedback. Long training phases, sometimes in combination with a learning criterion or performance contingent payment, are used to ensure intensive learning and memorization of the exemplars by the participants (e.g., Bröder, Newell, & Platzer, 2010; Hoffmann et al., 2013; Wirebring et al., 2018). In the testing phase, participants then have to judge the criteria of new stimuli and of already-learned exemplars. For instance, in Study 2 of von Helversen and Rieskamp (2009), participants had to evaluate the quality of job candidates (i.e., the criterion) on a scale of 1 to 100. The fictitious job candidates varied on six different cues (e.g., knowledge of programming languages, C++ vs. Java, knowledge of foreign languages, French vs. Turkish, etc.). The training phase consisted of 20 blocks with eight trials each. In each trial, participants had to judge one of the eight job candidates (i.e., the exemplars). After each trial, the participants received feedback about the number of points this candidate should receive and how close their estimate had been. In the testing phase, participants then had to judge 30 job candidates twice. From these 30 candidates, 22 were new candidates and eight were exemplar candidates from the training phase.

The parameters of the model of interest are often estimated based on the data of the last training blocks (e.g., Hoffmann et al., 2013; Juslin et al., 2008). These estimated parameters are then used to predict the data of each participant in the testing phase to avoid overfitting. The goodness-of-fit is then determined, for instance, via the root-mean-squared error (RMSE) between the model prediction and the participants' actual judgments or the Bayesian Information Criterion (BIC, Schwartz, 1978). The goodness-of-fit criteria are then often used together with qualitative indices of extrapolation and interpolation (e.g., Bröder & Gräf, 2018; Juslin et al., 2003) to compare the exemplar model with other possible judgment-process models, such as a rule-based model (e.g., Juslin et al., 2003; Hoffmann et al., 2013). Qualitative indices of extrapolation and interpolation are a valuable addition to quantitative goodness-of-fit measure, since exemplar models cannot predict judgments for new objects that extend beyond the range of learned criterion values, whereas rule-based

models can. Hence, testing for extrapolation in human judgments can help to distinguish the processes.

Exemplar model used in multiple-cue judgment research

The exemplar model we use as an example in this paper is based on the context model of Medin and Schaffer (1978) extended to account for the continuous criterion in multiple-cue judgments (Juslin et al., 2003). This and similar exemplar models have been used in many studies of multiple-cue judgments, where it is assumed that judgments are based on the memory of previously encountered exemplars (e.g., Bröder & Gräf, 2018; Bröder, Gräf, & Kieslich, 2017; Juslin et al., 2003; Hoffmann et al., 2013; Hoffmann et al., 2014; Hoffmann, von Helversen, Weilbächer, & Rieskamp, 2018; Karlsson et al., 2008; Platzer & Bröder, 2013; von Helversen & Rieskamp, 2008; von Helversen, Mata, & Olsson, 2010; Wirebring et al., 2018). According to this model, when a judgment is made about a probe (i.e., a stimulus that has to be judged), the judge considers the similarity of the probe to all of the previously encountered exemplars. Similarity then acts as a weight on the stored criterion values. When applied to a continuous criterion in a multiple-cue judgment task, the stored criterion value of a similar exemplar in memory has more influence on the judged criterion value of the probe, whereas the criterion value of a dissimilar exemplar receives less weight (Juslin et al., 2003). The similarity between a probe and an exemplar is determined by feature overlap. An exemplar with large feature overlap is more similar to the probe and thus has more impact on the judgment.

Regarding the formal definition, the model is based on the similarity S between a probe and the exemplars. It is assumed that the probe serves as a retrieval cue, activating previously encountered exemplars in memory. The probe $\vec{p}$ and each exemplar $\vec{e}_j$ are represented by vectors of D binary cues $\in \{0, 1\}$. The similarity parameters s_i, $i = 1, ..., D$, are the only free parameters in this model, defined on the interval $[0, 1]$. They determine how strongly a mismatch of objects on cue i influences the similarity S that can vary between 0 and 1. For simplicity, we assume the s_i to be constant across cues, that is, $s_i = s$, for all s_i (e.g., Bröder & Gräf, 2018; Juslin & Persson, 2002; von Helversen & Rieskamp, 2008).[1] The similarity $S(\vec{p}, \vec{e})$ between $\vec{p}$ and one exemplar $\vec{e}_j$ is determined according to the similarity rule of the context model (Medin & Schaffer, 1978):

$$S(\vec{p}, \vec{e}_j) = \prod_{i=1}^{D} d_i \text{ with } d_i = \begin{cases} 1 \text{ if } p_i = e_i \\ s \text{ if } p_i \neq e_i \end{cases} \quad (1)$$

where D is the number of cues of each object. For binary cues this simplifies to:

$$S(\vec{p}, \vec{e}_j) = s^{D-m} \quad (2)$$

where m is the number of matching cues between $\vec{p}$ and $\vec{e}_j$. The judged criterion value c' of the probe $\vec{p}$ is then the average of all n exemplar criterion values $\vec{c}$ in memory, weighted by the similarity of the respective exemplar to the probe:

$$c' = \frac{\sum_{j=1}^{n} S(\vec{p}, \vec{e}_j) * c(\vec{e}_j)}{\sum_{j=1}^{n} S(\vec{p}, \vec{e}_j)} \quad (3)$$

where $c(\vec{e}_j)$ is the criterion value of exemplar j. Equation 3 is the extension of the context model (Medin & Schaffer, 1978) from binary to a continuous criterion as suggest by Juslin et al. (2003; see also Elliot & Anderson, 1995; Juslin & Persson, 2002). It involves many simplifying assumptions, such as not directly modeling the exemplar retrieval process (cf., the EBRW model of Nosofsky & Palmeri, 1997), assuming that all exemplars are used when making a judgment (cf., Nosofsky & Palmeri, 1997; Albrecht, Hoffmann, Pleskac, Rieskamp, & von Helversen, 2019), and that all exemplars, their cues, and their criterion values are remembered and recalled without error. However, a detailed modeling of the recall and retrieval process is not intended with this model as it is used in the multiple-cue judgment literature, since it is mainly used as a tool to classify rule- and exemplar-based processes of judgments.

The *s* parameter

The s parameter from the model above is often called similarity parameter, since from an analytical point of view, it controls the similarity between two exemplars (Medin & Schaffer, 1978, see the example below). Psychologically, the s parameter has been interpreted as an attention parameter, since the perceived similarity of two exemplars decreases, when more attention is paid to potential cue mismatches (Medin & Schaffer, 1978; see also Juslin et al., 2003; von Helversen & Rieskamp, 2009). However, the s parameter can also be seen as a continuous measure of memory discriminability, where high values indicate no discrimination between exemplars and very small values indicate a perfect discrimination between exemplars in memory. The memory discriminability of exemplars increases when exemplars become well-learned as their memory traces become more distinct, reducing the perceived similarity (Shiffrin, Clark, & Ratcliff, 1990; Kılıç, Criss, Malmberg, & Shiffrin, 2017). The results reported

[1] There is also empirical data showing that this simplified version also outperforms the more complex model with a separate s_i parameter for each cue i in predicting individuals behavior (von Helversen & Rieskamp, 2008).

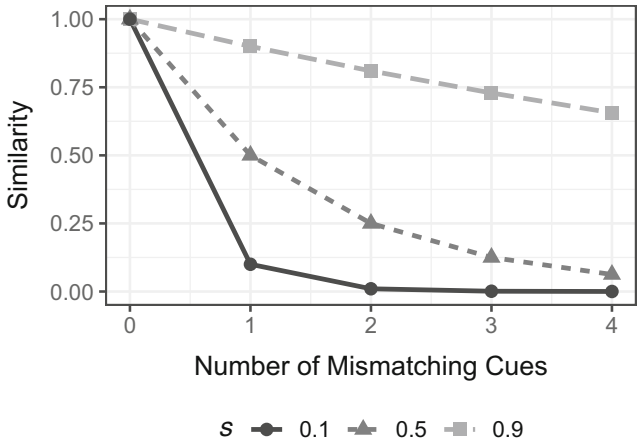


Fig. 1 The similarity between two stimuli for different numbers of mismatching cues and different values of s according to the context model (Medin & Schaffer, 1978)

in this article are relevant regardless of the preferred interpretation of s as either a memory or as an attention parameter.

To illustrate, Fig. 1 depicts the similarity between two hypothetical stimuli $\vec{a}$ and $\vec{b}$ with four cues each, for different numbers of mismatching cues (i.e., 0, 1, 2, 3, or 4), plotted for different values of s. For s = .9 (i.e., low discriminability), the similarity decreases rather slowly. However, for s = .1 (i.e., high discriminability), even a mismatch of only one cue (e.g., $\vec{a}$ = [1,1,1,1] and $\vec{b}$ = [0,1,1,1])) leads to a large decrease in similarity from 1 to 0.1. As more and more cues mismatch, the similarity asymptotically approaches 0. Due to the multiplicative combination of the mismatches (see Eq. 1), smaller values of s lead to a much steeper decrease of similarity with each mismatch and hence, much less influence of dissimilar exemplars on the judgment of the probe. This also implies that if s is equal, or very close to 0, only perfectly matching or very similar exemplars (if existent) will determine the judgment, otherwise, judgments are erratic. If s is equal, or very close to 1, every exemplar has the same influence, thus resulting in the prediction of the mean of the exemplar criterion values for each and every probe.

The psychological misspecification of the exemplar model in multiple-cue judgment research

The exemplar model as described above assumes that judging the criterion value of a probe *always* involves the reconstruction of the criterion value as a similarity-weighted average of all stored exemplar criterion values. This, however, seems psychologically implausible in typical multiple-cue judgment tasks where participants are repeatedly confronted with the same small set of judgment objects during the training phase. In this situation, we think that it is more realistic to assume that more and more exemplars become well learned, and when a probe is presented which is identical to one of the overlearned exemplars in memory, the criterion value of this very exemplar will be retrieved rather than building a similarity-weighted average of all exemplars. Hence, we assume that depending on the strength of an exemplar's memory representation, one of two qualitatively distinct processes will take place: If a strongly represented exemplar is found in memory which exactly matches the probe, the participant will simply recall this single exemplar's criterion value and report it as their judgment. We will henceforth refer to this process as *"(direct) recall"*. In the alternative case if there is no strong representation of a perfectly matching exemplar (the exemplar has not been overlearned, yet, or the probe is new), the similarity-weighted reconstruction as described in the original exemplar model takes place. Hence, all exemplars are recalled (cf., Nosofsky & Palmeri, 1997; Albrecht et al., 2019), and the judgment is built buy the similarity-weighted averaging process, for ease of differentiation, we will henceforth simply call this second process *"judgment"*. Of course, this judgment is also based on the recall of the exemplars. But it entails the additional process of a similarity-weighted integration of their criterion values in contrast to the "direct recall"-process mentioned above, which only entails the retrieval of just this one specific well-learned exemplar and then reproducing its retrieved criterion value. For ease of presentation, we use the shortcut terms "judgment" for the former integration process and "direct recall" for the latter simple recall process involving only the identical exemplar. Our conjecture is that lumping these qualitatively different processes together into one may severely distort the characterization of the process as well as the estimate of the similarity parameter s.

The reason for this conjecture is that the exemplar model can account for both types of processes (judgment and direct recall). In the judgment process, the response is a similarity-weighted average based on all exemplars. In the direct-recall process, it is only the identical exemplar that governs the response. Which of those two modes of aggregation is used depends on the s parameter and how the proposed psychological misspecification then might affect the estimation of the s parameter can be demonstrated with an example. Assume there are two exemplars and one to-be-judged stimulus with two binary cues as shown in Table 1. The to-be-judged stimulus is identical to Exemplar 1. Based on Eqs. 2 and 3, we can generate predictions for the unknown criterion value of the to-be-judged stimulus for a very high and a very low value of s. For $s = 1$ we get a prediction of $c' = 5$, which is the mean of the criterion values of the two exemplars. For $s = 0$ we get $c' = 3$, which is the criterion value of Exemplar 1. This implies that

Table 1 Cue and criterion values for two exemplars and one probe

	Cue 1	Cue 2	Criterion
Exemplar 1	0	1	3
Exemplar 2	1	1	7
Probe	0	1	**?**

Note. The probe is identical to the first exemplar

a very small s value leads to the prediction of the exact criterion value of the matching exemplar (if existent). Thus, when estimating the parameters from data by minimizing the distance between observed data and model-implied predictions, which can be seen as the reverse of prediction, s has to be as small as possible if two conditions are met: A probe is identical to one of the exemplars and the judged criterion of the probe is equal to the true criterion value of the matching exemplar. One instance where these conditions apply is when a participant has learned an exemplar and its respective criterion in an earlier stage of the experiment, and then later, when presented with the same exemplar again, recalls the learned criterion.

Therefore, we conjecture that the estimation of the s parameter is biased towards 0 if the responses of participants include direct recall of exemplars and their criterion values, and if all data points are jointly used to estimate the parameter. Furthermore, we predict that this problem is aggravated with increasing numbers of recalled exemplars, since there are more cases influencing the estimation of the s parameter. In addition, the model should show decreased model fit and make less accurate predictions when based on these biased parameter estimates.

In the present work, we show that disregarding the distinction between similarity-based judgment and direct recall leads to large errors in the estimation and impaired validity of model parameters. For this, we first present results from a computer simulation testing these predictions, showing bias in parameter estimation and how to avoid it. Next, we reanalyze data from five experiments and show the differences in parameter estimation and model fit between the usual procedure and a redefined procedure.[2]

Simulation

In this section, we show the severity of the problem and address the adequacy of a solution by running a computer simulation. The goal was to measure the bias in the estimation of the s parameter, for different true values of s and different recall probabilities P_r (i.e., the probability that the criterion value of an exemplar is recalled correctly). We compared three different ways for estimating the s parameter of the exemplar model presented above.

First, we used the typical procedure of multiple-cue judgment studies described above, which estimates the model parameters of the original exemplar model based on all data points regardless if it was a directly recalled exemplar, a not recalled exemplar, or a new stimulus. Based on the reasoning presented before, we expected that, when this $\hat{s}_{orig}$ parameter is estimated in this usual manner, it is more biased towards 0, the more correctly recalled exemplars there are in the data.

Second, since we propose that correctly recalled exemplars lead to a biased estimation of the s parameter, as a simple proof-of-concept, we estimated two different s parameters by splitting the data into two distinct sets of stimuli: Recalled exemplars (i.e., the well-learned, very distinct exemplars) versus not recalled exemplars and the new stimuli (i.e., less well-learned and less discriminable exemplars, as well as new stimuli). The $\hat{s}_{split}$ parameter, estimated only on the data set with not recalled exemplars and the new stimuli, should then be an unbiased estimator of s. However, this simple proof-of-concept is based on post hoc evaluation of the data (i.e., the classification in correctly recalled exemplar and other) and also reduces the amount of data used for estimating the parameter, since only a subset of the data is used. This method of splitting the data can be used as a heuristic remedy to arrive at appropriate estimates of s, however, if the extended model described next cannot be applied.

As a more elegant solution, we also used an extended version of the exemplar model which directly integrates the assumption that there are two distinct processes at work when people are confronted with already presented stimuli.[3] The graphical model is depicted in Fig. 2. This latent-mixture model (Zeigenfuse & Lee, 2010) assumes that the final response y_t of a participant in a trial t is generated by one of two possible processes, if the stimulus in this trial was part of the training phase: A direct retrieval of the learned criterion value of this trained exemplar c_t (= direct recall) or the similarity-weighted reconstruction

[2] All simulations and analyses were conducted using R (Version 4.0.2; R Core Team, 2020b) and the R-packages *afex* (Version 0.27.2; Singmann, Bolker, Westfall, Aust, & Ben-Shachar, 2020), *doSNOW* (Version 1.0.18; Corporation & Weston, 2019), *dplyr* (Version 1.0.0; Wickham, François, Henry, & Müller, 2020), *foreach* (Version 1.5.0; Microsoft & Weston, 2020), *foreign* (Version 0.8.80; R Core Team, 2020a), *ggplot2* (Version 3.3.2; Wickham, 2016), *lsr* (Version 0.5; Navarro, 2015), *MCMCvis* (Version 0.14.0; Youngflesh, 2018), *polspline* (Version 1.1.19; Kooperberg, 2020), *psych* (Version 1.9.12.31; Revelle, 2019), *purrr* (Version 0.3.4; Henry & Wickham, 2020), *Rcpp* (Version 1.0.5; Eddelbuettel & François, 2011; Eddelbuettel & Balamuta, 2017), *reshape2* (Version 1.4.4; Wickham, 2007), *runjags* (Version 2.0.4.6; Denwood, 2016), *tibble* (Version 3.0.1; Müller & Wickham, 2020), and *truncnorm* (Version 1.0.8; Mersmann, Trautmann, Steuer, & Bornkamp, 2018). The entire article was written with the *papaja*-package (Version 0.1.0.9997; Aust & Barth, 2020)

[3] We thank an anonymous Reviewer for this suggestion.

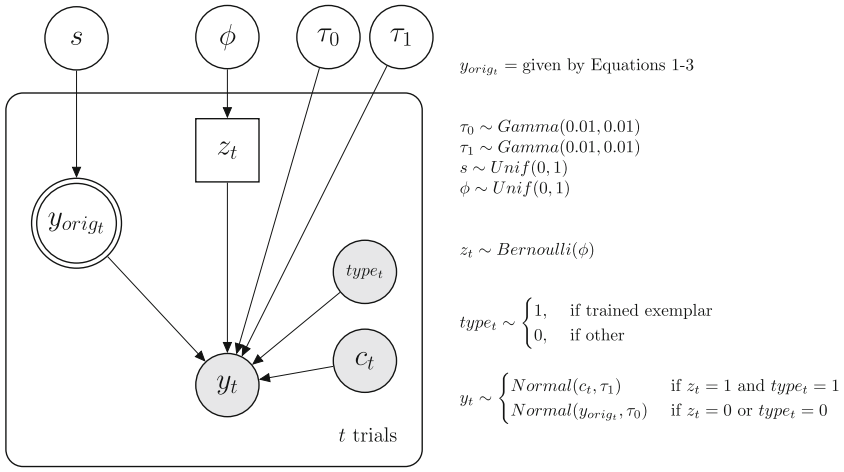


Fig. 2 Graphical model of the latent-mixture extension of the original exemplar model

as described in the original exemplar model y_{orig_t} (= judgment). Which data generating process is used, given the stimulus in this trial was a trained exemplar, is determined by an indicator variable z_t. If $z_t = 0$ the data y_t follow a normal distribution with precision τ_0 and centered around the prediction of the original exemplar model y_{orig_t}, which is based on the parameter s. If $z_t = 1$ the data y_t follow a normal distribution with precision τ_1 and centered around the learned criterion value of this exemplar c_t. The indicator z_t follows a Bernoulli distribution with parameter ϕ. This parameter ϕ represents the latent memory probability; this is the probability that a trained exemplar is directly recalled and the corresponding criterion value reproduced. To summarize, this extended latent-mixture model of the original exemplar model integrates the assumption that if a probe in a trial is a novel stimulus, the similarity-weighted average response based on the original exemplar model is used based on the parameter s. When the probe is a trained exemplar, the response is the directly recalled learned criterion value of this exemplar with probability ϕ and the similarity-weighted average response based on the original exemplar model with probability $1 - \phi$. The estimated $\hat{s}_{int}$ parameter should then also be unbiased, since the possibility of direct retrieval is already integrated into the model.

Procedure

In this simulation, we generated behavioral data by manipulating two independent variables (the true value of s and the probability that the criterion value of an exemplar is recalled correctly) and investigated how these variables influence the parameter estimation. A summary of the simulation procedure is shown in Algorithm 1 in the Appendix. In the first step of this simulation, we generated the stimulus matrix, consisting of 32 stimuli that can be created with five binary cues. The criterion values were computed according to a linear additive rule:

$$c = w_0 + cue_1 \times w_1 + cue_2 \times w_2 + cue_3 \times w_3 + cue_4 \times w_4 + cue_5 \times w_5. \quad (4)$$

where cue_i represents the binary cues and w_i the corresponding cue weights. Of the 32 stimuli, 12 are randomly selected as to-be-learned exemplars. In order to create realistic stimulus material used in actual experiments (e.g., Bröder et al., 2017; Bröder & Gräf, 2018), the four most extreme stimuli (i.e., the two stimuli with the highest and the two stimuli with the lowest criterion value) were never selected as exemplars. In addition, there was also a switch of criterion values between one pair of stimuli (i.e., if one stimulus a of this switch pair would have a criterion value of 31 and stimulus b of the pair a value of 59 , the new values after switching would be 59 for a and 31 for b). The cue weights w_i for cues $i = 0, ..., 5$ had to sum to 100 and were randomly drawn from a truncated normal distribution with $\mu = 20$, $\sigma = 10$, an upper bound of 100, and a lower bound of 0.

In the second step, we generated judgment data from this stimulus matrix according to Juslin et al.'s. (2003) version of the context model (Medin & Schaffer, 1978) presented above. The true s parameter varied in 4 steps from a very strict similarity criterion to a more lenient criterion, s = .001, .01, .3 or .8. The recall probability (P_r) could either be .1, .5 or 1. This means that, for instance for P_r = 0.5, there is a probability of .5 that an exemplar and its corresponding criterion value is recalled correctly and, therefore, there could be more or less than 50% of correctly recalled exemplars in a given iteration of the simulation when P_r = .5. A value of $Pr = 1$ indicated that every exemplar (and its criterion value) is recalled correctly and the judged criterion value of this exemplar is therefore its exact criterion value. A value of

$Pr = .1$ indicates that only very few exemplars are recalled exactly.[4] It should be noted that a recall probability of 1 is what most studies aim for when applying an extensive training phase. Also, as most exemplar models are based on the assumption that all exemplars and their corresponding criterion values are remembered correctly and are all used in the subsequent judgment process (cf., Nosofsky & Palmeri, 1997; Albrecht et al., 2019), participants should learn all exemplars correctly. Note also that we added no additional error to the generated judgment data, so in principle one would expect perfect parameter recovery.

In the third step, we estimated the $\hat{s}$ parameters with JAGS (Plummer, 2003) interfaced with R using the *runjags* package (Denwood, 2016), using each of the three methods. The results are based on MCMC chains with 5000 samples from each of two independent chains collected after 5000 burn-in samples were discarded, 5000 adaptive iterations, and thinning by recording every 5th sample. The convergence of the chains was checked by visual inspection and the standard $\hat{R}$ statistic (Brooks & Gelman, 1998).

In the final step, we computed the Bayes factors for model comparison between the original exemplar model ($\mathcal{M}_0$) and the latent-mixture model ($\mathcal{M}_1$). Since the original exemplar model is nested within the latent-mixture model when $\phi = 0$, we computed the Bayes factor based on the Savage–Dickey density ratio (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010; Vandekerckhove et al., 2015):

$$BF_{10} = \frac{p(\phi = 0|\mathcal{M}_1)}{p(\phi = 0|D, \mathcal{M}_1)} \tag{5}$$

where $p(\phi = 0|\mathcal{M}_1)$ is the density of the prior distribution of ϕ at 0 given $\mathcal{M}_1$, $p(\phi = 0|D, \mathcal{M}_1)$ is the density of the posterior distribution of ϕ at 0 given $\mathcal{M}_1$, and BF_{10} is the Bayes factor in favor of $\mathcal{M}_1$. The density of the posterior distribution was computed with the *dlogspline* function in the *polspline* package in R (Kooperberg, 2020). Since we used a uniform (0,1)-prior for ϕ, the density $p(\phi = x|\mathcal{M}_1)$ on any given point x is 1. The resulting Bayes factor BF_{10} then indicates the evidence of $\mathcal{M}_1$ compared to $\mathcal{M}_0$, or how much more probable the data are under $\mathcal{M}_1$ compared to $\mathcal{M}_0$ (Kass & Raftery, 1995; Morey, Romeijn, & Rouder, 2016; Vandekerckhove et al., 2015). For instance, a Bayes factor of BF_{10}=10 would indicate that the data are 10 times more likely to occur under $\mathcal{M}_1$ than under $\mathcal{M}_0$. In addition, we computed the root-mean-squared-error (RMSE) between the actual data and the median of the posterior predictive distribution in each trial of both models as an indicator for the prediction error margin of each model and since the RMSE is often used in multiple-cue judgment studies for model comparison (e.g., Hoffmann et al., 2013; Wirebring et al., 2018; von Helversen & Rieskamp, 2009).

All steps were repeated 200 times for each combination of true s parameter and P_r value, which leads to $200 \times 4 \times 3 = 2400$ simulated data sets in total. For each simulated data set, a new stimulus matrix, with different exemplars, cue weights, and criterion values was generated as described in the first step. The code of the simulation, the JAGS model codes, example of MCMC chains and $\hat{R}$ values of all three estimation methods for a randomly selected iteration of the simulation, and the results are available at the Open Science Framework (https://osf.io/b69f3/).

Results

Recovery *s*

The results of the simulation are displayed in Fig. 3 and Table 2. The first row in Fig. 3 shows that the recovered parameter $\hat{s}_{orig}$ of the original exemplar model was very close to the true s values, when P_r was small. However, with an increasing percentage of correctly recalled exemplars (P_r), $\hat{s}_{orig}$ increasingly deviated from s, with larger deviations for larger s values. For a high recall probability of $P_r = 1$, $\hat{s}_{orig}$ deviated strongly from the true s and was severely biased downward towards 0. In addition, $\hat{s}_{orig}$ was never larger than .17 for $s = .3$ and .27 for $s = .8$ when $P_r = 1$, which is less than half as large as the actual true value. Thus, the first row in Fig. 3 shows that the estimated $\hat{s}_{orig}$ parameter is a severely biased estimator of s if judgment and direct recall (or the data generated by these processes) are mixed. The bias of $\hat{s}_{orig}$ increases, when more exemplars are recalled directly. Yet, a good memory performance is exactly the goal researchers try to achieve, when they design their experiments with an extensive learning phase.

The second row of Fig. 3 shows the estimated $\hat{s}_{split}$ parameter, the parameter that was estimated only on the subset of the data without any directly recalled exemplars. When the parameter is estimated based only on new stimuli and criterion values that are not perfectly recalled, the recovered $\hat{s}_{split}$ values were identical to the true values, see again Table 2 for the descriptive values. Although not shown in Fig. 3, the estimated $\hat{s}_{split}$ parameter values based on the subset of only correctly recalled exemplars were mostly estimated close to 0, as to be expected.

The third row of Fig. 3 shows the recovered $\hat{s}_{int}$ parameter values from the latent-mixture extension of the original exemplar model. The results in Fig. 3 and displayed in Table 2 show that the true s parameter values are

[4] We used a value of .1 instead of 0, as this made it easier to ensure that there were at least two recalled exemplars, since it would lead to some problems later in the simulation when there were no or only one recalled exemplar when we estimated two separate s parameters on the two distinct subsets of data.

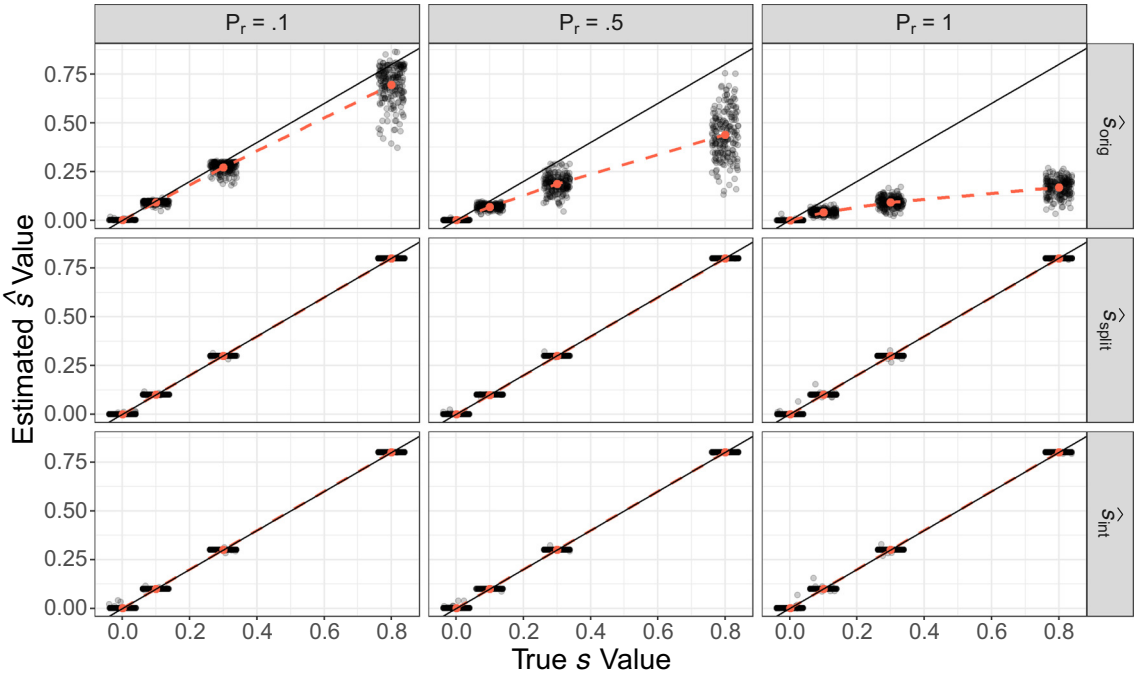


Fig. 3 Estimated s values for different true s values and different P_r for three different types of s parameter. The *black solid lines* represent what would be expected for perfect parameter recovery. *Red points* and *dashed lines* show and connect the means of 200 repetitions. $\hat{s}_{orig}$ is the estimated s parameter based on the original exemplar model. $\hat{s}_{split}$ is the estimated s parameter based on the original exemplar model, when only the subset of the data without any recalled exemplars was used. $\hat{s}_{int}$ is the estimated s parameter of the latent-mixture extension of the original exemplar model

recovered very well by $\hat{s}_{int}$ when the possibility of direct recall of trained exemplar was integrated into the model. In addition, we found that the ϕ parameter was very close to P_r on average over all iterations, except for s = .001, where the average ϕ was below .1, regardless of P_r (see Table S1 in the online supplementary material). We assume this is since the difference in criterion values between a directly recalled exemplar and the predicted value based on the exemplar model with s = .001 can be rather small (e.g., 47 and 47.002, see Table S2 in the online supplement) and the model then tends to classify directly recalled exemplars as "not-recalled" (i.e., $z_t = 0$).

Table 2 Means (and standard deviations) of the estimated s values for different true s values and different memory probabilities

		true s			
P_r	type	.001	.1	.3	.8
0.1	$\hat{s}_{orig}$	.001 (.002)	.092 (.008)	.272 (.026)	.696 (.096)
	$\hat{s}_{split}$	.001 (.002)	.100 (.001)	.300 (.002)	.800 (.000)
	$\hat{s}_{int}$	.001 (.004)	.100 (.002)	.300 (.002)	.800 (.000)
0.5	$\hat{s}_{orig}$	.001 (.001)	.069 (.012)	.186 (.045)	.439 (.121)
	$\hat{s}_{split}$	.001 (.002)	.100 (.001)	.300 (.002)	.800 (.000)
	$\hat{s}_{int}$	.001 (.004)	.100 (.001)	.300 (.002)	.800 (.000)
1.0	$\hat{s}_{orig}$	.001 (.002)	.042 (.013)	.093 (.027)	.169 (.044)
	$\hat{s}_{split}$	.002 (.005)	.100 (.004)	.300 (.003)	.800 (.001)
	$\hat{s}_{int}$	.001 (.005)	.100 (.004)	.300 (.003)	.800 (.001)

Note. P_r indicates the recall probability. Type indicates the type of s parameter: $\hat{s}_{orig}$ is the estimated s parameter based on the original exemplar model. $\hat{s}_{split}$ is the estimated s parameter based on the original exemplar model, when only the subset of the data without any recalled exemplars was used. $\hat{s}_{int}$ is the estimated s parameter of the latent-mixture extension of the original exemplar model

Table 3 Means (and standard deviations) of the $log(BF_{10})$

	true s			
P_r	.001	.1	.3	.8
0.1	− 2.32 (3.11)	2.92 (2.13)	3.06 (1.86)	3.13 (1.82)
0.5	− 1.83 (4.29)	8.74 (1.94)	8.76 (1.91)	8.71 (1.79)
1.0	− 2.30 (3.03)	18.45 (1.6)	18.55 (1.55)	18.49 (0.68)

Note. P_r indicates the recall probability

Model comparison

The Bayes factors as a mean of model comparison between the original exemplar model and the latent-mixture extension are shown in Table 3. When the data-generating s parameter was very small, the Bayes factor favors on average the original exemplar model, indicated by the negative $log(\text{BF}_{10})$, regardless of P_r. This is because the s parameter is with .001 already very close to 0, thus there is not much room for the downward biasing effect of correctly remembered exemplars and the Bayes factors then favors the less complex original exemplar model with fewer parameters. However, the average $log(\text{BF}_{10})$ increasingly favors the latent-mixture model with an increasing number of correctly recalled criterion values, with values more or less being constant for the different possible true s values.

We can find a similar pattern for the RMSE shown in Table 4. When the data were generated with a very small s value, we get a similar average low RMSE for the original exemplar model and its latent-mixture extension. However, the RMSE of the original exemplar model increases the larger the s parameter and the number of correctly recalled exemplars become.

Discussion

We ran a simulation to investigate the potential bias in the estimation of the s parameter, when one does not differentiate between recalled exemplars and not recalled exemplars as well as new stimuli.

The results suggest that the estimation of the s parameter as well as predictions based on this estimation can be inaccurate, when the distinction between directly recalled exemplars and judgment is not taken into account. The deviation of $\hat{s}_{orig}$ from s was small when either the true s parameter was small, or when P_r was small, that is, when there were only very few directly recalled exemplars. However, we found large biases in estimation as well as in predictions when there was a medium to large recall probability and true s value. The results show that the estimated $\hat{s}_{orig}$ is biased downwards, when the true s parameter and the recall probability was large. However, this large recall probability, as stated before, is exactly the outcome many experimenters aim for when designing their experiments with extensive learning phases: Many studies implement a learning criterion that participants have to reach to advance to the next phase of the experiment or terminate the learning phase before the maximal number of learning blocks (e.g., Bröder et al., 2010; Hoffmann et al., 2013; Wirebring et al., 2018). In addition, the number of learning blocks (i.e., the number of times an exemplar is presented) can range from 4 (Pachur & Olsson, 2012) up to 40 blocks (Wirebring, Stillesjö, Eriksson, Juslin, & Nyberg, 2018), with most studies using 8–10 blocks. Participants are instructed, and with these extensive learning phases also able, to memorize these stimuli and their respective criterion

Table 4 Means (and standard deviations) of the RMSE of the original exemplar model (orig) and the latent-mixture extension with integrated direct recall process (int)

		true s			
P_r	Type	.001	.1	.3	.8
0.1	$RMSE_{\mathcal{M}_0}$	0.03 (0.17)	0.70 (0.44)	1.67 (0.92)	3.36 (1.7)
	$RMSE_{\mathcal{M}_1}$	0.03 (0.16)	0.04 (0.25)	0.02 (0.14)	0 (0.01)
0.5	$RMSE_{\mathcal{M}_0}$	0.08 (0.35)	1.25 (0.33)	3.17 (0.75)	6.06 (1.29)
	$RMSE_{\mathcal{M}_1}$	0.08 (0.35)	0.02 (0.14)	0.01 (0.06)	0 (0.01)
1.0	$RMSE_{\mathcal{M}_0}$	0.04 (0.22)	1.40 (0.36)	3.23 (0.68)	6.13 (1.17)
	$RMSE_{\mathcal{M}_1}$	0.04 (0.21)	0.04 (0.23)	0.02 (0.11)	0 (0.01)

Note. P_r indicates the recall probability. $\mathcal{M}_0$ is the original exemplar model and $\mathcal{M}_1$ the latent-mixture extension

values. For instance, in Experiment 1B in Bröder, Gräf, and Kieslich (2017) participants showed an average correct recall rate (what we called here the recall probability P_r) of .79 ($SD = .23$) and 46.67% had a correct recall rate of .90 or more. Furthermore, it is not the recall probability per se, but the relative number of correct exemplars to all trials, that is, how many data points from all possible trials are correctly recalled exemplars, which drives this effect. The more recalled exemplars there are in the data, the stronger $\hat{s}_{orig}$ is biased towards 0. For instance, in our simulation, with 32 stimuli, 12 exemplars, and $P_r = 1$, there were $\frac{12}{32} = 37.50\%$ correctly remembered exemplars. The fact that parameters are often estimated on the data of the training blocks (e.g., Bröder & Gräf, 2018; Juslin et al., 2003; von Helversen & Rieskamp, 2008), where *all* stimuli are exemplars, makes this finding even more alarming. These findings could in principle explain why many studies find small $\hat{s}_{orig}$ values, since based on the results of the simulation, small $\hat{s}_{orig}$ values can arise both, from small true s values, but also from larger true s values, when combined with the often-achieved high number of correctly recalled exemplars in the data.

The results show that this bias in $\hat{s}_{orig}$ is caused by correctly remembered exemplars, this is instances where the judgment of the criterion value of a trained exemplar is identical to its true criterion value, since the bias disappears when the $\hat{s}_{split}$ was estimated only on the subset of the data without any recalled exemplars. As a more elegant solution, using the here presented latent-mixture extension of the original exemplar model where the possibility of correctly recalling a trained exemplar is integrated into the model also lead to an unbiased estimation of the s parameter. In fact, the few deviations of the estimated $\hat{s}_{int}$ from the underlying s parameter are probably due to the random simulation procedure and instances of unfortunate selections of exemplars and combinations of generated criterion values, which would not be used in a real experimental setting.

In the next section, we investigate if these effects reported here are likely to be found in real experimental data as well, by reanalyzing existing data from five different multiple-cue judgment experiments.

Re-analysis

We reanalyzed data from five different experiments from Bröder et al. (2017), Bröder and Gräf (2018), and one unpublished data set from the same lab group. The aim was to investigate if the effects found in the simulation extend to empirical data as well. Before we describe our general approach, we first outline the experimental procedure used in one of the experiments. The material and procedure in the other experiments were very similar to the one described below and can be found in the corresponding papers, or, for the unpublished data set, in the supplemental material. The code and results are again available at the Open Science Framework (https://osf.io/b69f3/).

Materials and procedure of the reanalyzed data sets

In Experiment 1A of Bröder et al. (2017), participants had to judge stimuli on a scale from 0 to 100 based on a set of four binary symptoms (e.g., fever vs. hypothermia), resulting in 16 different stimuli. They either had to judge the severity of a patient's disease or the toxicity of a bug. Since cue patterns and criterion values of both stimulus sets were identical and for reasons of simplicity, we will not make a distinction between the content domain in the subsequent analysis. The experiment itself consisted of three phases: a training phase, a decision phase, and a final testing phase. In the training phase, participants had to judge severity of illness of eight patients or the toxicity of eight bugs (the exemplars) and feedback about the actual criterion value was provided. Participants were instructed to either use the feedback about the correct criterion values to learn a mathematical rule connecting cue and criterion values (rule condition) or to memorize the objects and their values (exemplar condition). The training phase consisted of eight blocks with eight trials each (one for each exemplar). In the decision phase, participants had to choose the stimulus with the higher criterion value of 45 pairs of objects. These data, however, are not important for the current project. In the testing phase, participants had to judge the criterion values of all 16 stimuli (i.e., exemplars as well as new stimuli). They were instructed to either apply the mathematical rule they learned earlier (rule condition) or judge untrained objects by their similarity to the memorized objects (exemplar condition).

Method

Because of the often-documented "rule-bias" and since we were interested in the exemplar model, we chose the conditions of the five experiments in which exemplar-based processing was expected (or shown) to be most prevalent. We selected the data from the corresponding exemplar condition from each experiment, where participants were either directly instructed to use an exemplar-based approach (e.g., Experiment 1A in Bröder et al., 2017), or where an exemplar-based strategy was induced by experimental design (e.g., Bröder and Gräf, 2018). For instance, in Bröder and Gräf (2018), we only used the data from the condition where a dimensional cue format was combined with memory-based judgments, as more exemplar-based reasoning has been observed under these conditions

(Bröder, Newell, & Platzer, 2010; Platzer & Bröder, 2013). See Table 5 for a short overview of all experiments and the selected conditions.

We then again used JAGS (Plummer, 2003) to fit the original exemplar model and the latent-mixture model depicted in Fig. 2 to the data of the judgment phase of each experiment. We ran two MCMC chains with 5000 samples each with thinning by recording every 5th sample, after 15,000 burn-in samples and 15,000 adaptive iterations. The convergence of the chains was checked by visual inspection and the standard $\hat{R}$ statistic (Brooks & Gelman, 1998). We also computed Bayes factors for model comparison between the original exemplar model ($\mathcal{M}_0$) and the latent-mixture model ($\mathcal{M}_1$), using the Savage–Dickey density ratio (Wagenmakers et al., 2010). In addition, we computed the RMSE between the actual data and the median of the posterior predictive distribution of each model.

Hypotheses

We had three predictions based on the simulation results reported before. First, regarding the s parameter, we expected to find higher values for $\hat{s}_{int}$ than for $\hat{s}_{orig}$, since $\hat{s}_{orig}$ should be biased towards 0 when there are correctly recalled exemplars in the data (H1). Second, since $\hat{s}_{orig}$ becomes smaller on average when there are more correctly recalled exemplars, but $\hat{s}_{int}$ does not depend on the number of correctly recalled exemplars (see Fig. 3), we expected to find a negative correlation between $\hat{s}_{orig}$ and the number of correctly recalled exemplars. We also expected to find no such correlation with $\hat{s}_{int}$ and the number of correctly recalled exemplars (H2). Third, we expected to find that the data are better predicted by the latent-mixture model than by the original exemplar model, as indicated by a positive $log(\text{BF}_{10})$ (H3).

Results

H1. Differences in *s* Parameter

We conducted one-sided paired-samples t-tests for the differences between $\hat{s}_{int}$ and $\hat{s}_{orig}$ for each experiment. The results, together with the descriptive values, are shown in Table 6. Overall, we found significant differences between $\hat{s}_{int}$ and $\hat{s}_{orig}$, with all ps $< .001$ and ds ≥ 1.02. The median posterior estimates of $\hat{s}_{int}$ and $\hat{s}_{orig}$ for each person and each experiment depicted in Fig. 4 show that as hypothesized $\hat{s}_{int}$ is larger than $\hat{s}_{orig}$ for almost all participants. Indeed, there is only one instance were $\hat{s}_{int}$ (0.881) was smaller than $\hat{s}_{orig}$ (0.884) and this is for a participant who did not recall any exemplar correctly (i.e., $P_r = 0$). In addition, there was a very high correlation between the estimated latent memory probability parameter ϕ of the latent-mixture model and the empirical proportion of correctly recalled exemplars of each participant (Table 6), which supports the validity of the ϕ parameter.

H2. Correlations

To test our additional predictions we calculated the correlation between $\hat{s}_{int}$ as well as $\hat{s}_{orig}$ and the number of correctly recalled exemplars across participants and then

Table 5 Sample size, mean (and standard deviation) of proportion of correctly recalled exemplars, names and short description of the selected condition of each experiment

Exp.	Label	n	P_r	Selected condition	Short description of condition
Bröder et al. (2017) - 1A	171A	62	.43 (.32)	Exemplar instruction	Participants were instructed to use an exemplar-based strategy.
Bröder et al. (2017) - 1B	171B	30	.85 (.15)	Exemplar instruction	Participants were instructed to use an exemplar-based strategy.
Bröder et al. (2017) - 2	172	30	.69 (.31)	With picture	Each exemplar was always accompanied by a picture of a male person to facilitate exemplar-based processing.
Bröder and Gräf (2018)	18	30	.78 (.28)	Memory-based dimensions	A dimensional cue format was combined with memory-based judgments, to facilitate more exemplar-based reasoning.
Bröder and Gräf (unpublished)	XX	35	.54 (.30)	Long learning phase	Participants had a longer training phase to facilitate exemplar storage and thus exemplar-based processing.

Note. Label indicates the respective abbreviations used in subsequent tables and figures

n represents the respective sample size. P_r represents the proportion of correctly recalled exemplars

Table 6 Means (and standard deviations) for different s parameters, test statistics and effect sizes of the difference between $\hat{s}_{int}$ and $\hat{s}_{orig}$ for five data sets. Means (and standard deviations) of the latent memory parameter ϕ and its correlation with the empirical proportion of correctly recalled exemplars

					s parameters					
Exp.	n	P_r	ϕ	$r_{P_r \times \phi}$	$\hat{s}_{orig}$	$\hat{s}_{int}$	t	df	p	d
171A	62	.43 (.32)	.45 (.28)	.93 [.88,.96]	.34 (.23)	.48 (.21)	8.00	61	< .001	1.02
171B	30	.85 (.15)	.79 (.13)	1.00 [1.00,1.00]	.23 (.13)	.53 (.20)	12.88	29	< .001	2.35
172	30	.69 (.31)	.66 (.26)	1.00 [1.00,1.00]	.23 (.17)	.41 (.18)	10.79	29	< .001	1.97
18	30	.78 (.28)	.74 (.24)	1.00 [1.00,1.00]	.24 (.16)	.53 (.22)	9.36	29	< .001	1.71
XX	35	.54 (.30)	.53 (.25)	1.00 [1.00,1.00]	.36 (.24)	.53 (.23)	8.33	34	< .001	1.41

Note. n represents the respective sample size. P_r represents the proportion of correctly recalled exemplars

compared these two correlations with the test proposed by Dunn and Clark for the difference between two overlapping correlations based on dependent groups (Dunn & Clark, 1969; Hittner, May, & Silver, 2003). The results are shown in Table 7. In every data set, we found a stronger negative correlation of $\hat{s}_{orig}$ (range: -.55 to -.75) with the number of correctly recalled exemplars than for $\hat{s}_{int}$ (range: -.12 to -.47), with differences ranging from .20 to .48, $ps \leq .001$. Furthermore, as evident from Fig. 5, we found a similar pattern as in the simulation where there seems to be an upper bound for $\hat{s}_{orig}$ for high numbers of correctly recalled exemplars (see Fig. 3). Also evident from Fig. 5 is that, other than expected, there were two instances where $\hat{s}_{int}$ was still significantly related to the number of recalled exemplars.

However, as these are only correlational findings, the data from the unpublished experiment allowed us to address this prediction (i.e., that the number of correctly recalled exemplars affects $\hat{s}_{orig}$ but not $\hat{s}_{int}$) experimentally. This experiment consisted of two conditions which only differed in the length of the training phase (4 vs. 8 blocks, see the supplemental material for a more detailed description). This difference in length of the training phase should lead to a lower number of correctly learned exemplars for participants with a shorter training phase. This difference in the number of correctly learned exemplars should then lead to lower $\hat{s}_{orig}$ values when participants had a longer training phase and thus recalled more exemplars correctly, but it should not affect $\hat{s}_{int}$. As expected, participants with a shorter training phase recalled fewer exemplars in the final testing phase correctly ($M = 2.29, SD = 1.82$) than participants with a longer training phase ($M = 4.29, SD = 2.38$), $t(63.66) = -3.94$, $p < .001$, $d = 0.94$.

In addition, consistent with the previous results, the difference between $\hat{s}_{orig}$ and $\hat{s}_{int}$ was larger for participants with eight training blocks than for participants with only four training blocks ($F(1, 68) = 7.72$, $MSE = 0.01$, $p = .007$, $\hat{\eta}^2_G = .006$), as $\hat{s}_{orig}$ was lower in the long training condition ($M = 0.36, SD = 0.24$) than in the

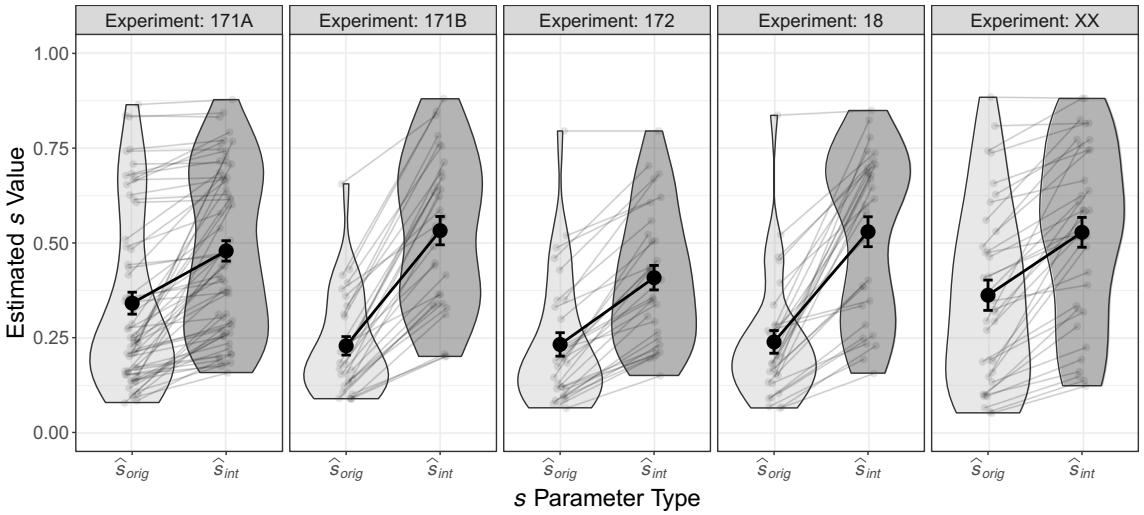


Fig. 4 Median posterior values of $\hat{s}_{int}$ and $\hat{s}_{orig}$ for each participant and for each data set. Black dots represent the means and the corresponding standard errors

Table 7 Correlations [and 95% CI] of $\hat{s}_{int}$ and $\hat{s}_{orig}$ with the number of correctly recalled exemplars and the test statistics regarding their differences in each data set

			r				
Exp.	n	P_r	$\hat{s}_{orig}$	$\hat{s}_{int}$	Δ	t	p
171A	62	.43 (.32)	− .60 [− .74,− .41]	− .15 [− .39,.10]	.45	6.65	< .001
171B	30	.85 (.15)	− .55 [− .76,− .23]	− .12 [− .46,.25]	.43	3.88	< .001
172	30	.69 (.31)	− .75 [− .87,− .53]	− .45 [− .70,− .10]	.30	4.46	< .001
18	30	.78 (.28)	− .68 [− .83,− .42]	− .20 [− .52,.17]	.48	3.58	< .001
XX	35	.54 (.30)	− .68 [− .82,− .44]	− .47 [− .70,− .17]	.20	2.99	.001

Note. n represents the respective sample size. P_r represents the proportion of correctly recalled exemplars. Δ represents the difference between the correlations

short training condition ($M = 0.45$, $SD = 0.24$), but there was no difference for $\hat{s}_{int}$ ($M_{short} = 0.54$, $SD = 0.23$, $M_{long} = 0.53$, $SD = 0.23$).

H3. Model comparison

As expected, the latent-mixture model was on average better able to account for the data, as indicated by the high positive $log(\text{BF}_{10})$, ranging from $M_{log(\text{BF}_{10})} = 3.61$ in Experiment 171A to in Experiment 171B $M_{log(\text{BF}_{10})} = 9.42$, see Table 8 for the full results. As evident from Fig. 6, there is some variation in the extent to which the latent-mixture model is better able to account for the data of individual participants, with even some instances where the original exemplar model was better able to predict their data. However, Fig. 6 does also show that these differences are mostly due to the difference in the proportion of correctly recalled exemplars of the participants ($r_{log(\text{BF}_{10})\times P_r} = .97$, $t(185) = 52.80$, $p < .001$). Furthermore, we found significant differences between the RMSE of the both models, with the latent-mixture model having a lower RMSE on average, $ps \leq .002$, $ds \geq .56$.. However, as evident from Table 8, although we found the expected differences in all data sets, some differences were rather small, for instance in Experiment 2 of Bröder et al. (2017).

Discussion

We reanalyzed data from five different experiments to investigate if the effects found in the previous simulation extend to empirical data as well. Results showed large differences between $\hat{s}_{orig}$ and $\hat{s}_{int}$ in all five data sets: $\hat{s}_{orig}$ was estimated to be smaller than $\hat{s}_{int}$ in each data set, as was suggested by the simulation and theoretical reasoning. It is also notable that the higher the proportion of correctly recalled exemplars was, the larger was the

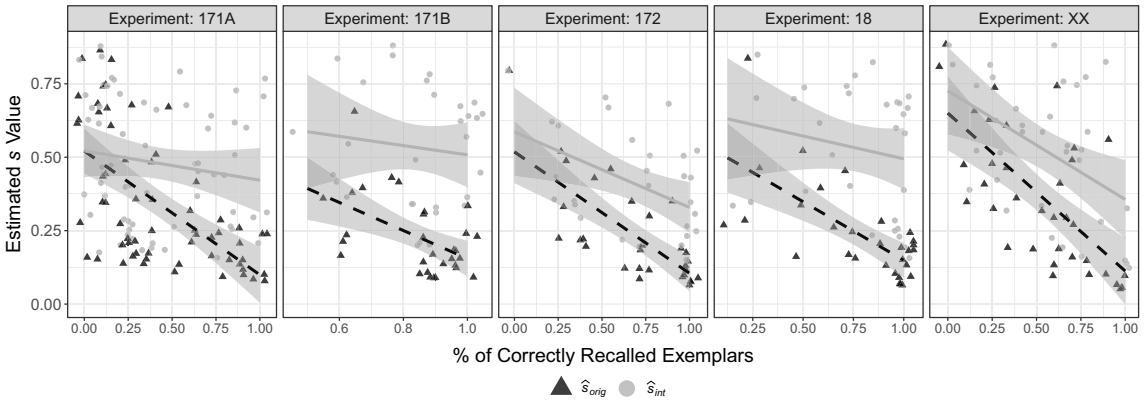


Fig. 5 Median posterior values of $\hat{s}_{int}$ (grey dots) and $\hat{s}_{orig}$ (black triangles) by proportion of correctly recalled exemplars, for each participant and for each data set. Lines and shaded areas represent the simple linear regression estimate and the 95% confidence interval

Table 8 Means (and standard deviations) of the RMSE of the original exemplar model or the latent-mixture model with integrated recall, with the corresponding test statistics and effect sizes of the difference between them, as well as the $log(\mathrm{BF}_{10})$, for five data sets

			$log(\mathrm{BF}_{10})$				$RMSE$					
Exp.	n	P_r	M	SD	Min	Max	$\mathcal{M}_0$	$\mathcal{M}_1$	t	df	p	d
171A	62	.43 (.32)	3.61	4.59	− 2.14	12.01	13.88 (3.61)	13.25 (3.59)	7.07	61	< .001	0.90
171B	30	.85 (.15)	9.42	1.58	6.99	12.91	11.62 (3.45)	10.40 (3.32)	5.51	29	< .001	1.01
172	30	.69 (.31)	7.30	3.84	− 2.12	12.04	12.56 (3.17)	12.19 (3.02)	3.08	29	.002	0.56
18	30	.78 (.28)	8.45	3.65	− 1.27	12.41	12.27 (3.30)	11.20 (3.23)	5.53	29	< .001	1.01
XX	35	.54 (.30)	5.63	4.20	− 2.17	10.92	13.73 (3.86)	12.86 (3.54)	4.99	34	< .001	0.84

Note. n represents the respective sample size. P_r represents the proportion of correctly recalled exemplars. $\mathcal{M}_0$ is the original exemplar model and $\mathcal{M}_1$ the latent-mixture extension

difference between $\hat{s}_{orig}$ and $\hat{s}_{int}$. Furthermore, correlational results on the participant level also showed that $\hat{s}_{orig}$ highly depends on the number of correctly recalled exemplars, with participants who recalled more exemplars correctly having lower $\hat{s}_{orig}$ values. Although not always independent from the number of recalled exemplars as originally expected, $\hat{s}_{int}$ was clearly less strongly correlated with the number of recalled exemplars. These results were further corroborated with experimental data of one data set showing that participants with a longer training phase recalled more exemplars correctly and had lower $\hat{s}_{orig}$ values than participants with a shorter training phase. Yet, there was no difference in $\hat{s}_{int}$. One might argue that it is plausible that participants who better learned the exemplars are better able to differentiate between them, which in turn is captured by lower s values in the model. Although this might be true, we still would argue that the simulation results presented before clearly show that the relationship between a higher number of correctly recalled exemplars and lower $\hat{s}_{orig}$ values can be a pure methodological and technical artifact. Taken together, these results would also suggest that the difference between $\hat{s}_{orig}$ and $\hat{s}_{int}$ would be even larger when parameters are estimated on data from the training phase (by $\hat{s}_{orig}$ becoming even smaller), since there are only trained exemplars in the learning phase and thus, the bias of $\hat{s}_{orig}$ can be even greater.

Moreover, we found that overall and for most individual participants, the latent-mixture model integrating a direct recall process of trained exemplars is better able to account for the data, where for participants with a very low number of correctly recalled exemplars the original exemplar model was preferred. However, although the Bayes factors give strong to extreme (overall) evidence for the latent-mixture model, the differences in RMSE of both models, which is often used in multiple cue judgment studies as a goodness-of-fit criterion, were rather small for some experiments, although we found somewhat larger differences in the simulation.

To further investigate this, we ran a simulation similar to the one described before, but with settings based on the experiments we reanalyzed. That is, we used the same stimuli, exemplars, and criterion values as in the studies we reanalyzed. In addition, in each of the 500 repetitions of the simulation we drew the recall probability P_r and

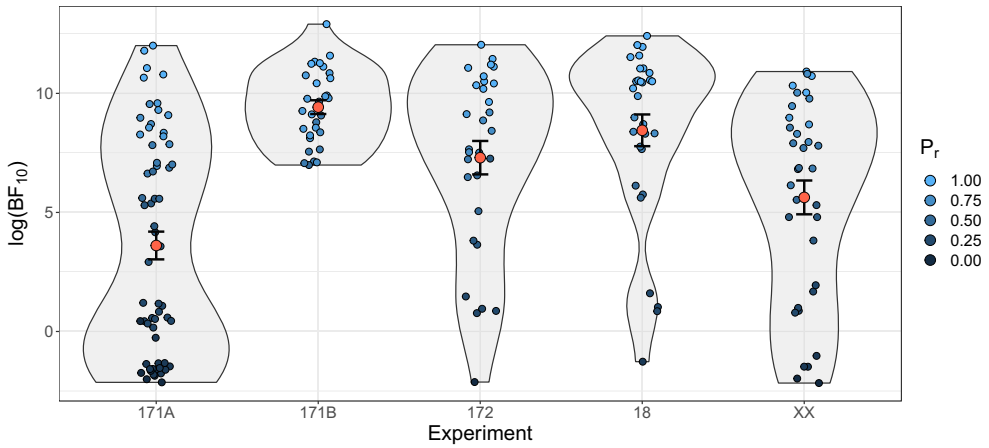


Fig. 6 The $log(\mathrm{BF}_{10})$ colored by the proportion of correctly recalled exemplars (P_r) for each participant and for each data set. The *red dots* represent the means and the corresponding standard errors

Table 9 Means (and standard deviations) of parameter estimates, RMSE, and effect sizes, from simulated as well as empirical data

	s parameters		RMSE			
Type	$\hat{s}_{orig}$	$\hat{s}_{int}$	$\mathcal{M}_0$	$\mathcal{M}_1$	d	$log(\text{BF}_{10})$
Empirical	.28 (.19)	.50 (.21)	12.81 (3.48)	11.98 (3.34)	0.86 (0.86)	6.88 (3.57)
Simulation	.32 (.19)	.52 (.20)	13.72 (4.62)	12.76 (4.82)	1.13 (0.27)	6.42 (3.45)

Note. d represents Cohen's d for a paired-sample t test. $\mathcal{M}_0$ is the original exemplar model and $\mathcal{M}_1$ the latent-mixture extension

the true s value from $beta$ distributions with similar means and standard deviations as found in the experiments we reanalyzed[5]. We then used these parameters and the stimuli to generate judgment data in each repetition according to the exemplar model presented in Eqs. 1 to 3 and then added normal distributed error with $\mu = 0$ and $\sigma \sim N(17, 6)$[6] with a lower bound of 0 and upper bound of 100. To be clear, we only defined the stimuli, the criterion values, s, and P_r, based on the data of the reanalyzed experiments. However, we did not define or set any constraints on the resulting RMSE. We then estimated the parameters and assessed the RMSE as in the simulation reported above. The results are shown in Table 9. As intended, the average s parameters over all simulations were similar to the average values found in the empirical experiments. Furthermore, we found that although the overall RMSE was a little bit higher in the simulation, the average effect sizes of the RMSE difference between both models and the average $log(\text{BF}_{10})$ were similar to the ones found in the empirical data sets. This suggests that the results we found regarding the differences in RMSE and $log(\text{BF}_{10})$ are somewhat typical for the specific memory performance observed in the studies and the specific stimuli and range of criterion values used in the experiments we reanalyzed.

A limitation of the results presented here is that the procedure of the experiments we reanalyzed were very similar to one another. Also, despite having different content domains, the stimuli used in all experiments (i.e., number of cues, number of stimuli, exemplars, and criterion values) were the same in all experiments. Therefore, it is still open to which extent the results generalize to other experiments, with different stimuli, cues, and exemplars.

General discussion

We proposed that in the typical experimental procedure in the multiple-cue judgment literature, the responses of participants are a mixture of two qualitatively distinct cognitive processes (similarity-based judgments and direct recall) and that disregarding this distinction can lead to biased estimation and impaired validity of parameters. We ran a simulation and reanalyzed data from five experiments to investigate the properties and extents of this issue, as well as the adequacy of a solution. Results of the simulation and the reanalysis showed that the estimation of the s parameter of the context model (Medin & Schaffer, 1978) extended to account for the continuous criterion in multiple-cue judgments (Juslin et al., 2003) can be severely biased towards 0 and that the model fit decreases if one does not differentiate between recalled exemplars and other stimuli, especially for larger values of the underlying s parameter and if more exemplars are recalled correctly. Furthermore, we found that on an individual level, the usually estimated $\hat{s}_{orig}$ parameter was very strongly negatively correlated with the number of correctly recalled exemplars in all five data sets, whereas the redefined parameter $\hat{s}_{int}$ showed a weaker to no relationship. The simulation and the reanalyzed data sets showed that the predictive performance of the exemplar model is impaired when one does not differentiate between recalled exemplars and other stimuli.

These findings have several implications. First, we showed that the standard procedure for estimating the s parameter can lead to biased parameter estimates and impaired fit of the model. However, this is not a problem with the model itself. The problem is rather the adaptation of the experimental design from categorization research which involves having few overlearned stimuli (e.g., Medin & Schaffer, 1978; Nosofsky & Palmeri, 1998), to multiple-cue judgment research, in order to apply the exemplar models also to multiple-cue judgments (Juslin et al., 2003). The important difference between categorization and judgment is the scale of the criterion. In categorization research the criterion is categorical, for instance, two categories A or B (e.g., Medin & Schaffer, 1978; Juslin et al., 2003; Smith & Minda, 1998). In this case, multiple exemplars share the same criterion value, since there are several exemplars in category A and several exemplars in category B. Thus, there is no unique exemplar-criterion-value combination as in the multiple-cue judgment literature, were most exemplars have their unique criterion value. This combination of very

[5]We used the means and standard deviations of the unbiased $\hat{s}_{int}$ estimates to define the distribution of true s values in this simulation.

[6]These values were chosen randomly and are not based on empirical data.

few well learned exemplars with their unique criterion values leads to the biased estimation of the *s* parameter we presented here. We would thus propose that the bias of the *s* parameter is less profound in a categorization experiment, where multiple exemplars share the same criterion. In addition, there are also other paradigms were the bias of the *s* parameter should be not necessarily a problem. For instance, if participants get no direct feedback about the criterion value, they are not able to just learn and recall the exact criterion value (e.g., Pachur & Olsson, 2012). Also, there are studies in the multiple-cue judgment and in the categorization literature, were stimuli are often defined by continuous dimensions such as length, size, and brightness rather than by binary features (e.g., Brehmer, 1972; Nosofsky & Alfonso-Reese, 1999; Ratcliff & Rouder, 1998), which leads to a large set of unique stimuli and exemplars, which also makes it harder for participants to learn specific exemplars and their criterion values. However, on a psychological level, the mixture between different process still is a problem in these cases.

Second, the findings presented in this work could explain why previous studies found rather small values for the *s* parameter of the exemplar model. For instance, von Helversen and Rieskamp (2008) found average estimated parameter values between .001 and .17 (according to von Helversen & Rieskamp, 2009), Juslin, Karlsson, and Olsson (2008) found average values from .14 to .36, and Bröder and Gräf (2018) found an average *s* value of .11. As evident from Fig. 3, when participants recalled most of the exemplars the estimated parameter becomes rather small, even when the true underlying *s* value was large. In the simulation, there was an upper bound of .27 for the estimated $\hat{s}_{orig}$ parameter when $s = .8$ and when the recall of exemplars was perfect.

Third, because of the biased estimation of the model parameter, the goodness-of-fit and predictive performance of the model are impaired. But having non-biased parameter estimates becomes important since indices of model fit and model comparison (e.g., RMSE, BIC, BF) are often used to classify participants as users of a rule-based or an exemplar-based process (e.g., Hoffmann et al., 2013; von Helversen & Rieskamp, 2008; Wirebring et al., 2018). For example, von Helversen and Rieskamp (2008) estimated the parameters of different candidate models (e.g., the exemplar model introduced here) by minimizing the RMSE for participants' judgments in the last three blocks of the training phase. They then compared the RMSE between the model predictions and the actual data in the test phase to determine which process participants relied on. The predictions for the test phase were based on the estimated parameters of the training phase. By neglecting the different retrieval-based processes and estimating only one distorted *s* parameter, the exemplar model may suffer an undeserved disadvantage in the model comparisons which in the end could even result in an overestimation of rule-based processes in judgment. However, this problem may be less severe in studies comparing rule-based and exemplar-based models by qualitative indices of extrapolation and interpolation (e.g., Bröder & Gräf, 2018; Juslin et al., 2003), which are arguably less sensitive to the exact value of the *s* parameter and thus less affected by the results reported in this work.

One possible solution, which we presented here in the paper, is the latent-mixture extension of the original exemplar model shown in Fig. 2. As demonstrated in the simulation, the integration of the possibility of direct recall of learned exemplars ensures a valid estimation of the parameter of interest. Furthermore, this latent-mixture model is generally preferred over the original exemplar model, when participants remembered at least some exemplars correctly. However, so far, the model assumes a very simple and error-free direct retrieval process of the criterion value of a learned exemplar in a trial, where the corresponding criterion value of the exemplar is always correctly remembered, for example, there is no confusion between similar exemplars. In addition, there might be other possible solutions, such as splitting the data into correctly remembered exemplars and other stimuli, as demonstrated in the simulation. Although this was our initial idea of fixing this issue, this approach has several disadvantages over the latent-mixture approach. For instance, the split-solution is based on the post-hoc evaluation of the observed data, where the data is divided into two different sets (recalled exemplars vs. not recalled exemplars and new stimuli) and model parameters are then estimated separately for each set. In this dichotomization procedure, it would also be an approximation to categorize all exactly remembered exemplars in the set representing the pure recall process and all other trials in the second set, and then estimating one overall *s* parameter for each set. Furthermore, the latent-mixture model models the underlying psychological processed explicitly. Another possible solution would be to either do not give participants feedback about the exact criterion value (e.g., Pachur & Olsson, 2012) or to include some exemplars in the training phase for which no feedback about the criterion value is given, similar to Experiment 2 of Bröder et al. (2017), and then estimate the *s* parameter only on these exemplars.

Limitations

There are some limitations to this work. First, throughout this article, we focused on Medin and Schaffer's (1978) context model extended to continuous judgment (Juslin & Persson, 2002; Juslin et al., 2003) as an exemplar model. However, several multiple-cue judgment studies

(e.g., Hoffmann et al., 2014; Scheibehenne & Pachur, 2015; Pachur & Olsson, 2012) use another exemplar model, the GCM of Nosofsky (1984). We conjecture that the results found here for the s parameter of the context model extend to the sensitivity parameter c (also sometimes denoted as h) of the GCM as well, since the context model (Medin & Schaffer, 1978) is a special case of the generalized context model (Nosofsky, 1984) and the s parameter of the context model is related to the sensitivity parameter c through a monotonic function (see the supplemental materials). A second limitation is that for reasons of simplicity, we did not manipulate or randomize some factors in the simulation such as the general form of the criterion value function (e.g., linear, cubic, exponential), the number of cues, or the dimensionality of cues (binary vs. non-binary). However, since the biased estimation of the s parameter is due to having judgments of exemplars identical to the criterion value of the exemplar, irrelevant of how the criterion value is computed or how many cues there are, we expect the effects would be the same. Third, the data sets we reanalyzed originate from one lab-group and used similar materials (i.e., cue patters, criterion value function, criterion values, and exemplars). Different experimental materials may differ in the magnitude of effects we reported here. However, we expect that since the effects stem from the combination of the paradigm, the unmodified transfer of the model to this paradigm, and the estimation procedure used in multiple-cue judgment studies, and that we found the same effects in the simulation which used somewhat different materials (randomized criterion value functions, criterion values, and exemplars), the results should generalize to other studies as well.

Conclusions

We showed that the paradigm commonly used in multiple-cue judgment research in combination with the way models are fitted to the data can lead to biased estimates and impaired validity of parameters, as well as negatively affect the fit of the models. Researchers should be aware of the different possible psychological processes underlying their data and incorporate it in their analysis or experimental design if necessary.

Appendix

Algorithm 1 Simulation procedure.

Input: $s \in \{.001, .01, .3, .8\}$; $P_r \in \{.01, .05, 1.0\}$
Output: $\hat{s}_{orig}, \hat{s}_{split}, \hat{s}_{int}, RMSE, BF_{10}$
for each s and P_r **do repeat** 200 times
    Generate a 32 x 5 stimulus matrix S from five binary cues ▷ Step 1
    Draw five cue weights and intercept $w_i \sim N(20, 10)T(0, 100)$; $\sum_{i=0}^{5} w_i \leq 100$
    Compute the criterion value c_i for each stimulus i Eq. 4
    Select 12 random stimuli from S as exemplars E
    Switch criterion values of two randomly drawn exemplars
    Generate judgments j_i for each stimulus i according to Eqs. 1 and 3 ▷ Step 2
    Replace j_i of exemplar i with c_i with probability P_r to generate j_i'
    Estimate parameters: ▷ Step 3
    - $\hat{s}_{orig}$ on all data points and with the original exemplar model
    - $\hat{s}_{split_1}$ on non-exemplars or not recalled exemplars (i.e., $j_i' \neq c_i$)
    - $\hat{s}_{split_0}$ on recalled exemplars (i.e., $j_i' = c_i$)
    - $\hat{s}_{int}$ on all data points and with the latent-mixture extended exemplar model
    BF_{10} & $RMSE$: ▷ Step 4
    - Calculate the BF_{10} via $\frac{p(\phi=0|\mathcal{M}_1)}{p(\phi=0|D,\mathcal{M}_1)}$
    - Calculate the RMSE
    return $\hat{s}_{orig}, \hat{s}_{split}, \hat{s}_{int}, RMSE, BF_{10}$
end for

Acknowledgements The authors thank Sophie Scharf, Stefan Radev, Benjamin Hilbig, and Martin Schnuerch for helpful discussions and comments on an earlier version of the manuscript. This research was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG, GRK 2277) to the research training group Statistical Modeling in Psychology (SMiP).

Funding Open Access funding enabled and organized by Projekt DEAL.

Open practices statement R scripts and results for all simulations and analyses, as well as the simulation data are available at the Open Science Framework (https://osf.io/b69f3/). Raw data of the reanalyzed experiments are available upon request from the second author. The original experiments were not preregistered.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommonshorg/licenses/by/4.0/.

References

Albrecht, R., Hoffmann, J., Pleskac, T., 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*. https://doi.org/10.1037/xlm0000772

Aust, F., & Barth, M. (2020). papaja: create APA manuscripts with R markdown. R package version 0.1.0.9997. Retrieved from https://github.com/crsh/papaja

Brehmer, B. (1972). Cue utilization and cue consistency in multiple-cue probability learning. *Organizational Behavior and Human Performance*, *8*(2), 286–296.

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.

Bröder, A., & Gräf, M. (2018). Retrieval from memory and cue complexity both trigger exemplar-based processes in judgment. *Journal of Cognitive Psychology*, *30*(4), 406–417. https://doi.org/10.1080/20445911.2018.1444613

Bröder, A., Newell, B. R., & Platzer, C. (2010). Cue integration vs. exemplar-based reasoning in multi-attribute decisions from memory: a matter of cue representation. *Judgment and Decision Making*, *5*(5), 326–338.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455.

Corporation, M., & Weston, S. (2019). doSNOW: foreach parallel adaptor for the 'snow' package. R package version 1.0.18. Retrieved from https://CRAN.R-project.org/package=doSNOW

Denwood, M. J. (2016). runjags: an R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, *71*(9), 1–25. https://doi.org/10.18637/jss.v071.i09

Dunn, O. J., & Clark, V. (1969). Correlation coefficients measured on the same individuals. *Journal of the American Statistical Association*, *64*(325), 366–377. https://doi.org/10.1080/01621459.1969.10500981

Eddelbuettel, D., & Balamuta, J. J. (2017). Extending extitR with extitC++: a brief introduction to extitRcpp. *PeerJ Preprints*, *5*, e3188v1. https://doi.org/10.7287/peerj.preprints.3188v1

Eddelbuettel, D., & François, R. (2011). Rcpp: seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. https://doi.org/10.18637/jss.v040.i08

Elliott, S. W., & Anderson, J. R. (1995). Effect of memory decay on predictions from changing categories. *Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 815–836.

Henry, L., & Wickham, H. (2020). Purrr: functional programming tools. R package version 0.3.4. Retrieved from https://CRAN.R-project.org/package=purrr

Hintzman, D. L. (1984). MINERVA 2: a simulation model of human memory. *Behavior Research Methods, Instruments, and Computers*, *16*(2), 96–101. https://doi.org/10.3758/BF03202365

Hittner, J. B., May, K., & Silver, N. C. (2003). A Monte Carlo evaluation of tests for comparing dependent correlations. *Journal of General Psychology*, *130*(2), 149–168. https://doi.org/10.1080/00221300309601282

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2013). Deliberation's blindsight: how cognitive load can improve judgments. *Psychological Science*, *24*(6), 869–879. https://doi.org/10.1177/0956797612463581

Hoffmann, J. A., von Helversen, B., Weilbächer, R. A., & Rieskamp, J. (2018). Tracing the path of forgetting in rule abstraction and exemplar retrieval. *Quarterly Journal of Experimental Psychology*, *71*(11), 2261–2281. https://doi.org/10.1177/1747021817739861

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

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: a division of labor hypothesis. *Cognition*, *106*(1), 259–298. https://doi.org/10.1016/j.cognition.2007.02.003

Juslin, P., Olsson, H., & Olsson, A. C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology*, *132*(1), 133–156. https://doi.org/10.1037/0096-3445.132.1.133

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): a lazy algorithm for probabilistic inference from generic knowledge. *Cognitive Science*, *26*(5), 563–607. https://doi.org/10.1016/S0364-0213(02)00083-6

Karlsson, L., Juslin, P., & Olsson, H. (2008). Exemplar-based inference in multi-attribute decision making: contingent, not automatic, strategy shifts? *Judgment and Decision Making*, *3*(3), 244–260.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kılıç, A., Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2017). Models that allow us to perceive the world more accurately also allow us to remember past events more accurately via differentiation. *Cognitive Psychology*, *92*, 65–86. https://doi.org/10.1016/j.cogpsych.2016.11.005

Kooperberg, C. (2020). Polspline: polynomial spline routines. R package version 1.1.19. Retrieved from https://CRAN.R-project.org/package=polspline

Mata, R., von Helversen, B., Karlsson, L., & Cüpper, L. (2012). Adult age differences in categorization and multiple-cue judgment. *Developmental Psychology*, *48*(4), 1188–1201. https://doi.org/10.1037/a0026084

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Mersmann, O., Trautmann, H., Steuer, D., & Bornkamp, B. (2018). Truncnorm: truncated normal distribution. R package version 1.0-8. Retrieved from https://CRAN.R-project.org/package=truncnorm

Microsoft, & Weston, S. (2020). Foreach: provides foreach looping construct. R package version 1.5.0. Retrieved from https://CRAN.R-project.org/package=foreach

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18. https://doi.org/10.1016/j.jmp.2015.11.001

Müller, K., & Wickham, H. (2020). Tibble: simple data frames. R package version 3.0.1. Retrieved from https://CRAN.R-project.org/package=tibble

Navarro, D. (2015). Learning statistics with r: a tutorial for psychology students and other beginners. (version 0.5). R package version 0.5 University of Adelaide. Adelaide, Australia. Retrieved from http://ua.edu.au/ccs/teaching/lsr

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104–114. https://doi.org/10.1037/0278-7393.10.1.104

Nosofsky, R. M., & Alfonso-Reese, L. A. (1999). Effects of similarity and practice on speeded classification response times and accuracies: further tests of an exemplar-retrieval model. *Memory and Cognition*, *27*(1), 78–93. https://doi.org/10.3758/BF03201215

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(2), 266–300. https://doi.org/10.1037/0033-295X.104.2.266

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

Pachur, T., & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology*, *65*(2), 207–240. https://doi.org/10.1016/j.cogpsych.2012.03.003

Persson, M., & Rieskamp, J. (2009). Inferences from memory: strategy-and exemplar-based judgment models compared. *Acta Psychologica*, *130*(1), 25–37. https://doi.org/10.1016/j.actpsy.2008.09.010

Platzer, C., & Bröder, A. (2013). When the rule is ruled out: exemplars and rules in decisions from memory. *Journal of Behavioral Decision Making*, *26*, 429–441. https://doi.org/10.1002/bdm

Plummer, M. (2003). JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*. Vienna, Austria, (Vol. 124, pp. 1–10).

R Core Team (2020a). Foreign: read data stored by minitab, s, sas, spss, stata, systat, weka, dbase,... R package version 0.8-80. Retrieved from https://CRAN.R-project.org/package=foreign

R Core Team (2020b). R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356. https://doi.org/10.1111/1467-9280.00067

Revelle, W. (2019). Psych: procedures for psychological, psychometric, and personality research. R package version 1.9.12. Northwestern University. Evanston, Illinois. Retrieved from https://CRAN.R-project.org/package=psych

Scheibehenne, B., & Pachur, T. (2015). Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice. *Psychonomic Bulletin and Review*, *22*(2), 391–407. https://doi.org/10.3758/s13423-014-0684-4

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

Shiffrin, R. M., Clark, S. E., & Ratcliff, R. (1990). List-strength effect: II. theoretical mechanisms. *Journal of Experimental psychology: Learning, Memory, and Cognition*, *16*(2), 179–195.

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2020). Afex: analysis of factorial experiments. R package version 0.27-2. Retrieved from https://CRAN.R-project.org/package=afex

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: the early epochs of category learning. *Journal of Experimental psychology: Learning, Memory, and Cognition*, *24*(6), 1411.

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.) *Oxford library of psychology. The Oxford handbook of computational and mathematical psychology*, (pp. 300–319): Oxford University Press.

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*(1), 220–229. https://doi.org/10.1037/a0016690

von Helversen, B., & Rieskamp, J. (2008). The mapping model: a cognitive theory of quantitative estimation. *Journal of Experimental Psychology*, *137*(1), 73–96. https://doi.org/10.1037/0096-3445.137.1.73

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*(4), 867–889. https://doi.org/10.1037/a0015501

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: a tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), 1–20. Retrieved from http://www.jstatsoft.org/v21/i12/

Wickham, H. (2016). *Ggplot2: elegant graphics for data analysis*. New York: Springer. Retrieved from https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., & Müller, K. (2020). Dplyr: a grammar of data manipulation. R package version 1.0.0. Retrieved from https://CRAN.R-project.org/package=dplyr

Wirebring, L. K., Stillesjö, S., Eriksson, J., Juslin, P., & Nyberg, L. (2018). A similarity-based process for human judgment in the parietal cortex. *Frontiers in Human Neuroscience*, *12*, 1–18. https://doi.org/10.3389/fnhum.2018.00481

Youngflesh, C. (2018). Mcmcvis: tools to visualize, manipulate, and summarize mcmc output. *Journal of Open Source Software*, *3*(24), 640. https://doi.org/10.21105/joss.00640

Zeigenfuse, M. D., & Lee, M. D. (2010). A general latent assignment approach for modeling psychological contaminants. *Journal of Mathematical Psychology*, *54*(4), 352–362. https://doi.org/10.1016/j.jmp.2010.04.001

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

What is the airspeed velocity of an unladen swallow? Modeling numerical judgments of realisitc stimuli.

Izydorczyk, D. & Bröder, A.

The manuscript can be found under:

https://psyarxiv.com/3avwk

This dissertation was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG) to the Research Training Group "Statistical Modeling in Psychology" (GRK 2277).

Cover image created by David Izydorczyk & Judith Wieland using DALL-E 2 and GarryKillian/Freepik as background.