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

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

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

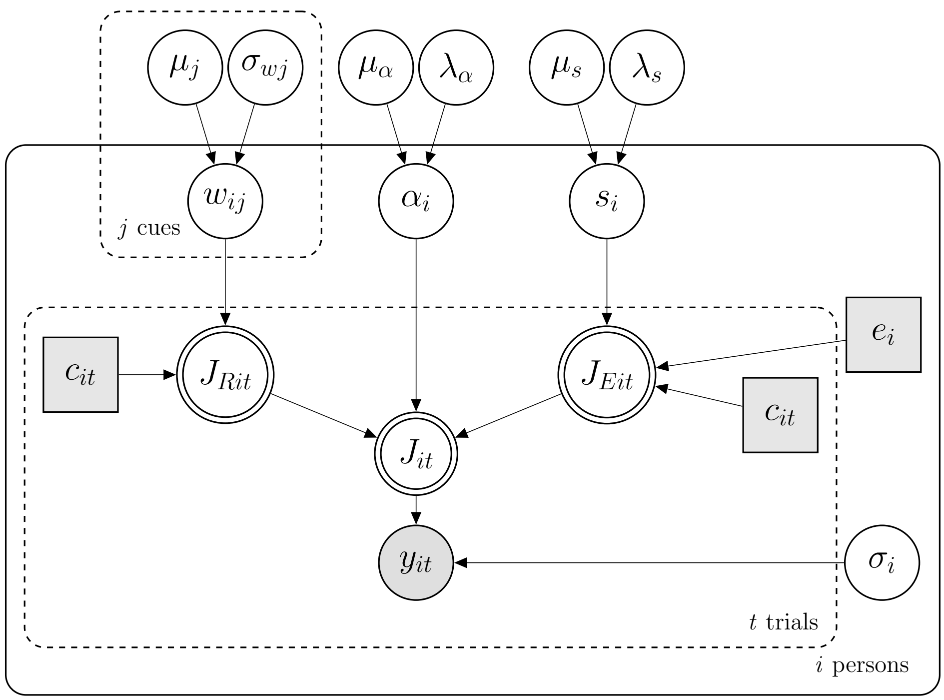
[6] The number of parameters of a model is only one factor determining the complexity of the model. Other factors such as the parameter range and the functional form (i.e., how the parameters are combined) also influence a model's complexity.

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:

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

where cue_j represents the binary cues coded with 0 and 1 and $w_{j_{\mathrm{gen}}}$ 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.

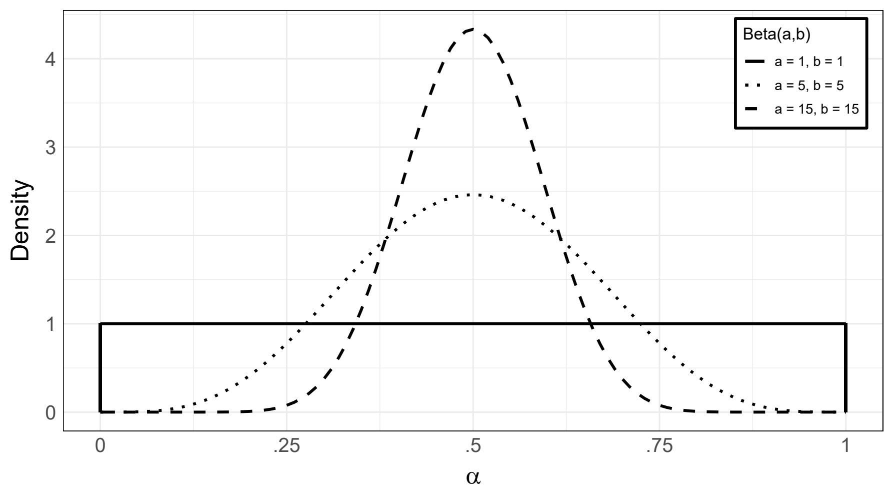
In the second step, we drew the generating parameter values for $n = 30$ simulated participants in the first simulation, which is a typical sample size in such experiments (e.g., Bröder et al., 2017; Hoffmann et al., 2013; Trippas & Pachur, 2019)), and $n = 50$ in the second simulation. In the first 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

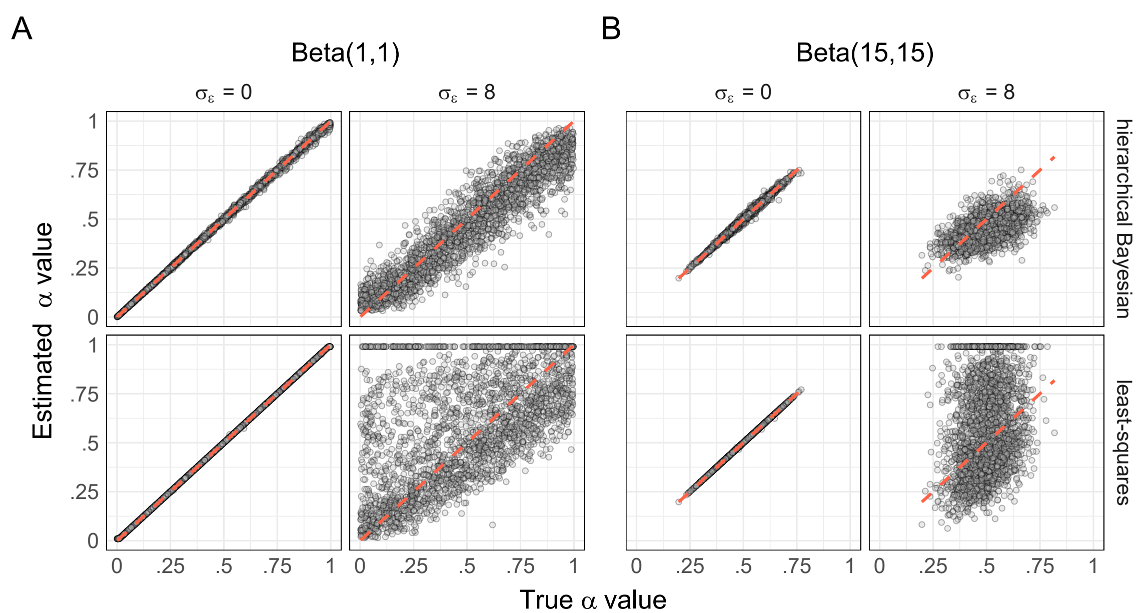
The results for the second simulations with the largest amount of noise ($\sigma_\epsilon = 8$) are shown in Table 1. The full results can be found in the online materials of this project.

[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*

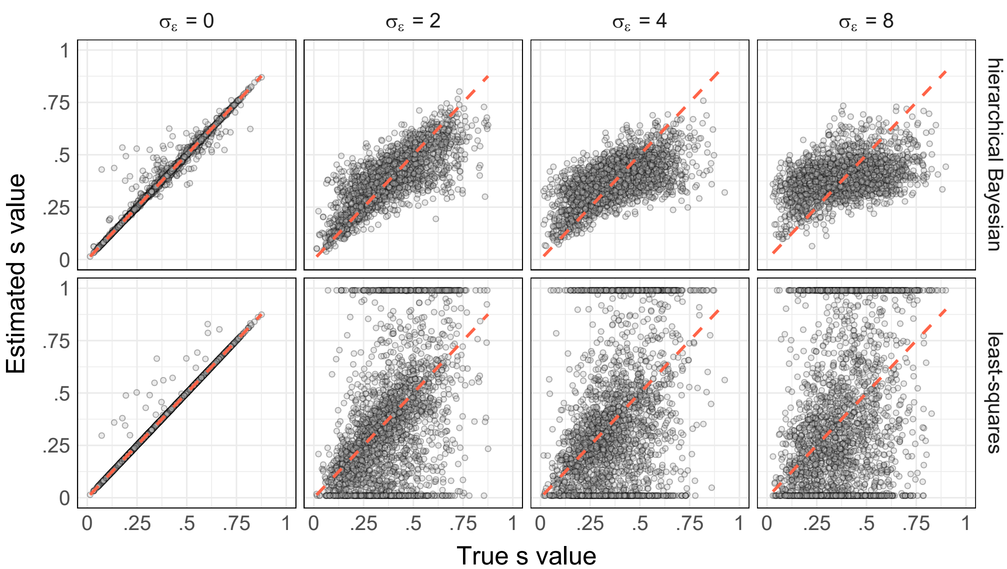


Note. The true α parameter values were drawn from a **A** Beta(1,1) or **B** Beta(15,15) distribution and for either no ($\sigma_\epsilon = 0$) or large ($\sigma_\epsilon = 8$) amounts of noise in the generated data. The two rows correspond to the different parameter estimation methods.

***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).

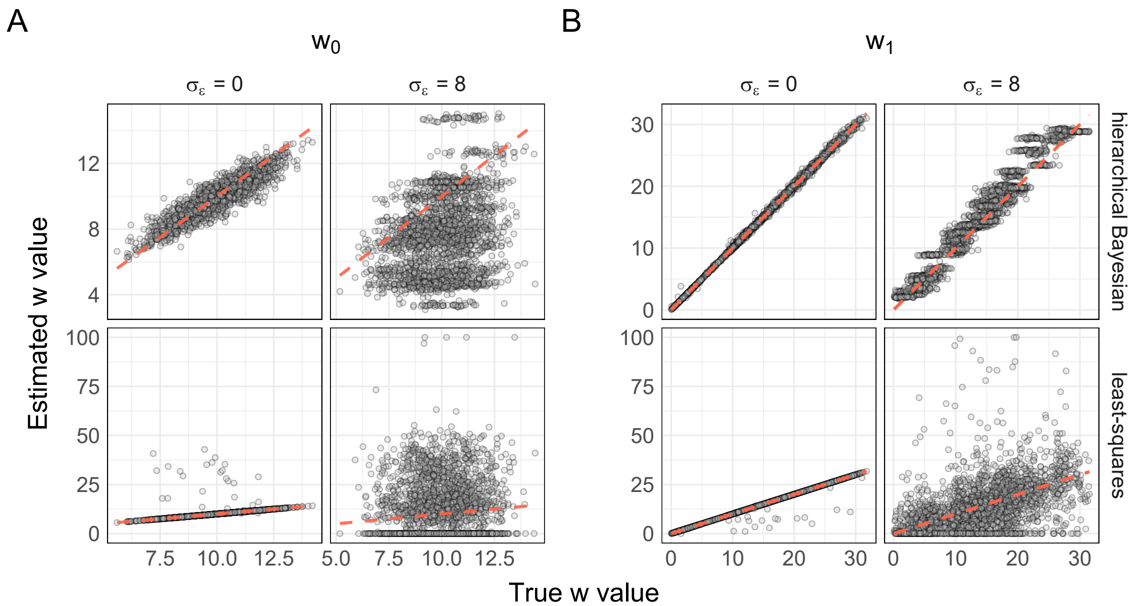


Note. The α parameter values were drawn from a Beta(1,1) distribution. The two rows correspond to the different parameter estimation methods.

***w_j*.** 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 (σ_ϵ = 0) or large (σ_ϵ = 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 (σ_ϵ = 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 $\alpha \sim$ 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 $r = .76$.

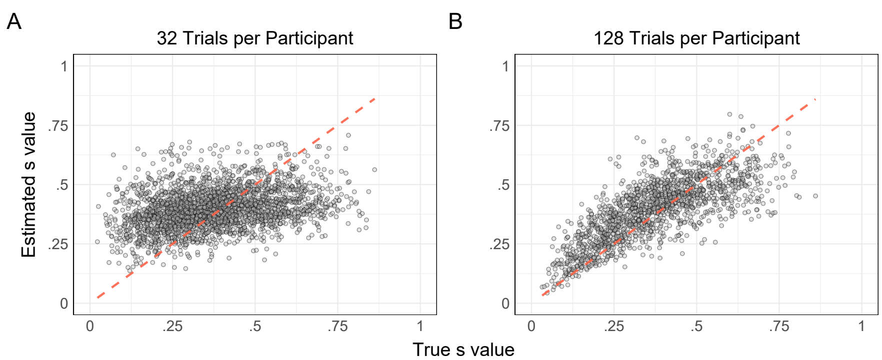
It should also be noted that, although we report here the results for all individual level

---

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)

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

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

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

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

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

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

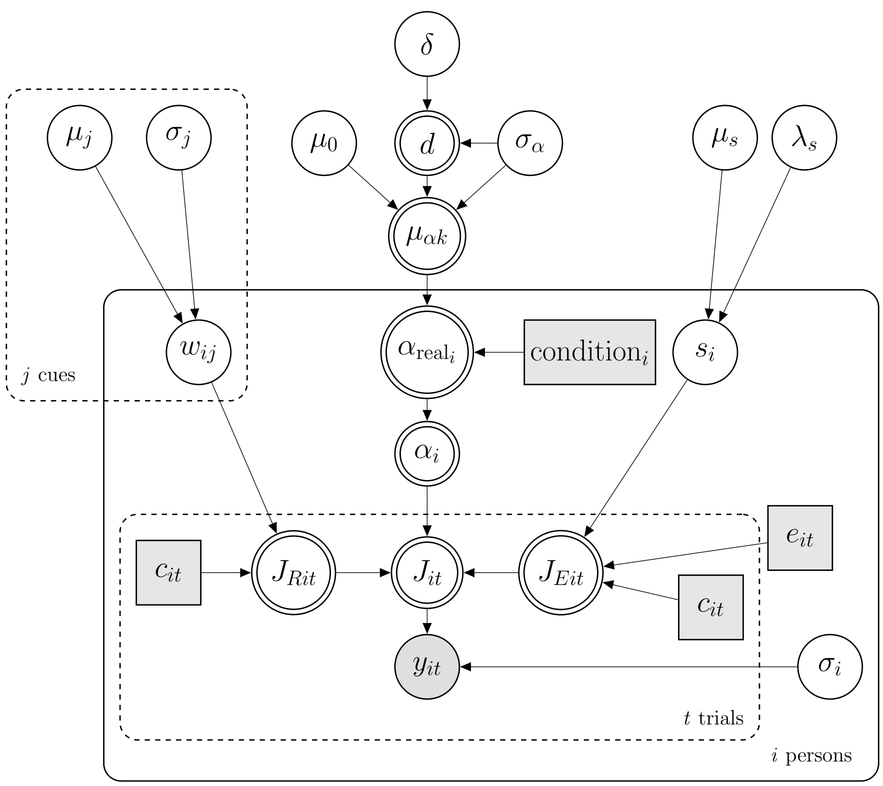
$$\tau_\alpha = \frac{1}{\sigma_\alpha^2} \tag{13}$$

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

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.

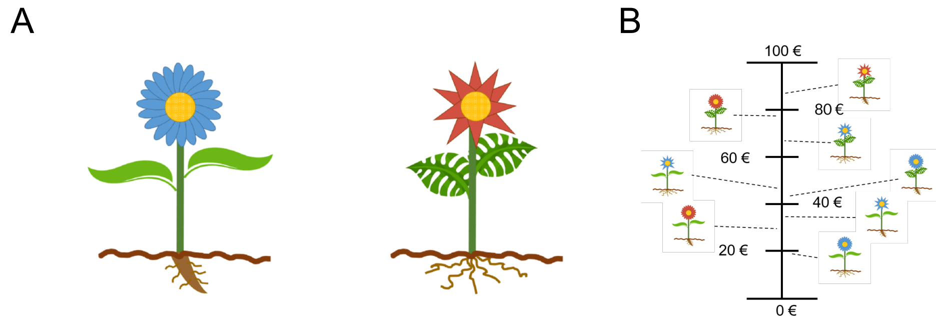
Thus, instead of instructing participants on what to learn during a learning phase as in Bröder et al. (2017) Exp. 1A (i.e., the criterion values, the cues, or a rule connecting both), we directly gave participants the information they should have learned to respond with a exemplar-based or rule-based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

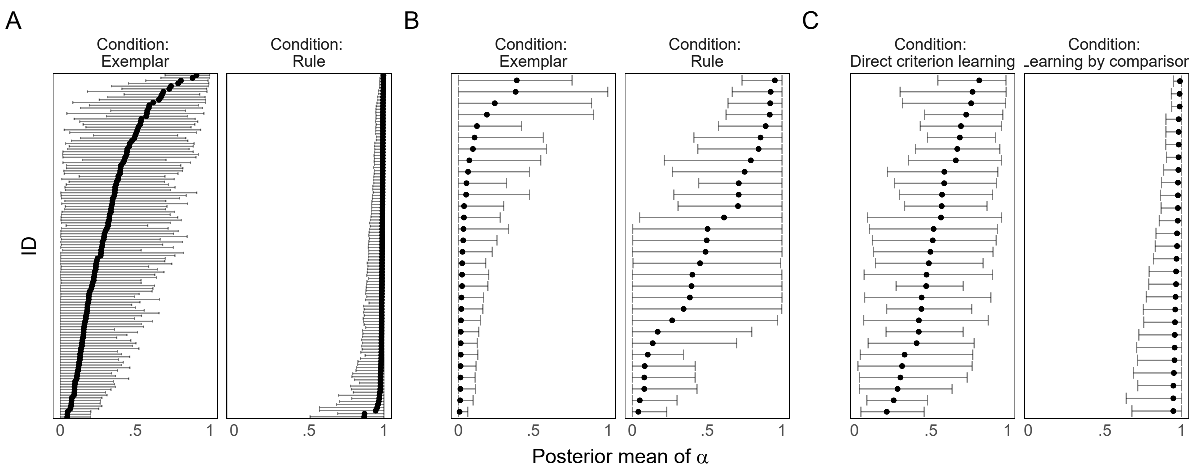
[14] We initially wanted to collect participants only via university mailing lists, however, due to very slow recruitment because to the COVID Pandemic we decided to also recruit participants via Prolific Academic Ltd.

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 $BF_{10} > 1000$ times

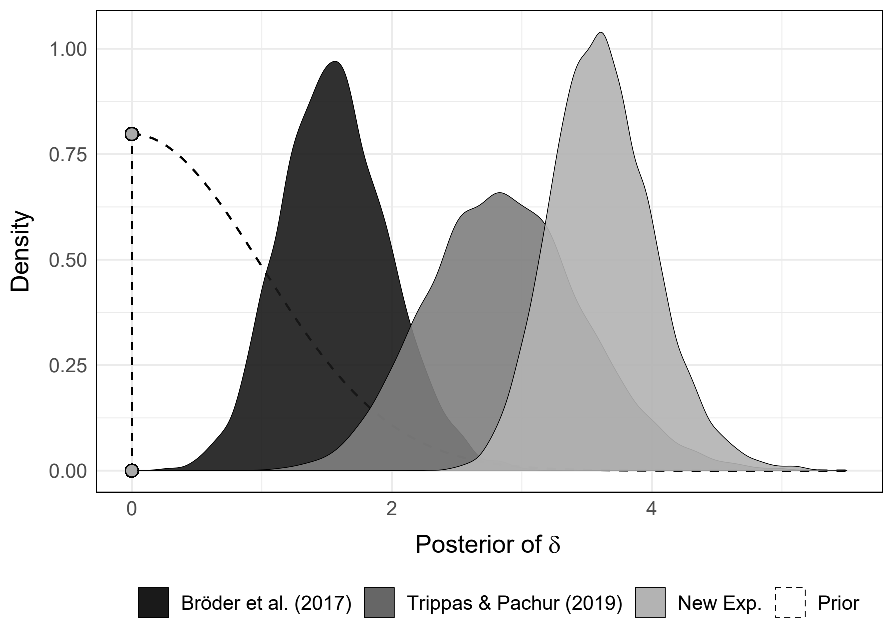
[15] We did not preregister the last two filtering steps (i.e., based on language and RMSE). However, the results presented in this section do not change substantially, when the excluded participants were included.

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 $\text{BF}_{10} = 367.15$ 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

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