Judgment and Decision Making, Vol. 12, No. 5, September 2017, pp. 491–506

Measuring the relative contributions of rule-based and exemplar-based processes in judgment: Validation of a simple model

Arndt Bröder* Michael Gräf† Pascal J. Kieslich‡

Abstract

Judgments and decisions can rely on rules to integrate cue information or on the retrieval of similar exemplars from memory. Research on exemplar-based processes in judgment has discovered several task variables influencing the dominant mode of processing. This research often aggregates data across participants or classifies them as using either exemplar-based or cue-based processing. It has been argued for theoretical and empirical reasons that both kinds of processes might operate together or in parallel. Hence, a classification of strategies may be a severe oversimplification that also sacrifices statistical power to detect task effects. We present a simple measurement tool combining both processing modes. The simple model contains a mixture parameter quantifying the relative contribution of both kinds of processes in a judgment and decision task. In three experiments, we validate the measurement model by demonstrating that instructions and task variables affect the mixture parameter in predictable ways, both in memory-based and screen-based judgments.

Keywords: judgment; exemplar models; measurement

1 Introduction

In research on categorization and judgment, at least two different classes of inference processes have been proposed for combining the probabilistic cues that inform about category membership or the quantitative value of a distal variable: 1., cue information is combined according to some *rule* (or heuristic), such as a linear weighted integration or a lexicographic consultation of cues (Brehmer, 1994; Gigerenzer & Goldstein, 1996); or, 2., the inference made is based on the general similarity of the to-be-judged probe to a set of *exemplars* stored in long-term memory (e.g., Medin & Schaffer, 1978; Nosofsky, 1984). Whereas the validity of exemplar models for categorization has been explored rather extensively, the application to quantitative judgment is relatively recent, sparked by the seminal work of Juslin and colleagues (e.g., Juslin & Persson, 2002; Juslin, Olsson & Olsson, 2003). Both kinds of processes rely on different knowledge representations: Whereas rule-based reasoning requires some knowledge about the covariation of individual cues with the judgment criterion ("cue abstraction"), exemplar-based inference relies on stored exemplars in long-term memory without any need for abstraction. Since judgments are based on the global similarity of feature patterns between probes and exemplars, there is no necessity to know the directions or validities of individual cues. However, this saved cognitive pre-processing effort comes with the prize of reduced reliability compared to rule-based judgments.

There have now been several demonstrations that both kinds of processes are recruited for numerical judgments. Recently, Hoffmann, von Helversen, and Rieskamp (2014) further validated the distinction of both processes by showing in an ingenious large scale study that they draw on different cognitive resources, namely working memory and long-term memory for rule-based and exemplar-based processing, respectively. Generally, people appear to prefer rule-based processing, but when cue abstraction is difficult for variable reasons, they switch to exemplar-based processes (Bröder, Newell & Platzer, 2010; Hoffmann, von Helversen & Rieskamp, 2013; Juslin et al., 2003; Karlsson, Juslin & Olsson, 2008; Platzer & Bröder, 2013).

1.1 Theoretical and empirical reasons for dual route

The studies mentioned above treat exemplar-based and rule-based judgment as exclusive modes of thinking, at least at an empirical level, classifying participants as apparent users of one or the other process (e.g., Bröder et al., 2010; Pachur & Olsson, 2012; Persson & Rieskamp, 2009; Platzer & Bröder, 2013). This coarse-grained dichotomy, however, does not do justice to the theoretical development in which both processing modes are often viewed as complementary. For example, Juslin, Karlsson, and Olsson's (2008) SIGMA

This research was funded by grant BR 2130/12-1 of the Deutsche Forschungsgemeinschaft (DFG) awarded to the first author as part of the research unit FOR 1410 "Contextualized decision making".

*School of Social Sciences, University of Mannheim, D-68131 Mannheim, Germany. E-mail: broeder@uni-mannheim.de.

†University of Administrative Sciences, Speyer, Germany.

‡School of Social Sciences, University of Mannheim, Germany.

https://doi.org/10.1017/S1930297500006513 Published online by Cambridge University Press

model assumes both kinds of representations that can be used as input for a joint judgment mechanism. Although the authors often speak of "shifts" from one distinct process to the other (see also Karlsson et al., 2008), they acknowledge that "the detailed processes are likely to involve, at least to some extent, a mix between the processes, between and within participants" (Juslin et al., 2008, p. 291). However, the authors leave open whether processes can also be mixed within one judgment or only across different judgments.

In the categorization literature, various hybrid models postulating both rules and exemplar-based processes have been proposed. As one example, Erickson and Kruschke's (1998; Kruschke & Erickson, 1994) ATRIUM model assumes two modules for rule and exemplar representations, respectively. The rule model can handle simple one-dimensional rules, and the exemplar model is based on Kruschke's (Kruschke, 1992) ALCOVE model. A stochastic gating mechanism governs the probability with which one of the respective modules is activated in a trial to determine the response. In this vein, ATRIUM can handle rule-based classification as well as similarity-based influences as demonstrated in both experiments reported by Erickson and Kruschke (1998).

In fact, empirical evidence gathered in the last two decades suggests that category judgments are probably always influenced by similarity-based processes, even if rules are available. For example, Regehr and Brooks (1993, see also Brooks & Hannah, 2006; Hannah & Brooks, 2009; Thibaut & Gelaes, 2006) showed that although a perfectly predictable rule for category membership was present (and in some experiments explicitly communicated to the participants), perceptual similarity of features and/or exemplars affected classification speed and/or accuracy. Similarly, using less complex stimuli, Erickson and Kruschke (1998) showed that participants were able to extrapolate a rule to new transfer stimuli, but classification probabilities were nevertheless affected by the similarity to exception stimuli encountered during learning. Hence, exemplar-based processes in categorization appear to be ubiquitous. Recently, Hahn, Prat-Sala, Pothos, and Brumby (2010) argued that these classic demonstrations of similarity effects all involved situations with graded category structures in which exemplar-based processing was to some extent functional. In the four experiments reported by Hahn et al., manipulated similarity was entirely based on irrelevant features, there was a very simple, explicit, and perfectly predictive three-features rule (Exp. 1, 3 & 4) or an even simpler one-feature rule (Exp. 2), and categorization by similarity was detrimental to performance. Nevertheless, the authors demonstrated similarity effects on accuracy and/or response times in all four experiments. They argue that the influence of similarity is probably automatic and beyond strategic control. Von Helversen, Herzog, and Rieskamp (2014) transferred this to a situation of (hypothetical) personnel selection and demonstrated an apparently unavoidable influence of facial similarity between candidates on judgments of their competence.

To summarize, exemplar- or similarity-effects appear to be ubiquitous in category judgments even in the presence of simply applicable rules. In research on (numerical) judgment, exemplar-based reasoning has hitherto been viewed more as an exception or a "backup system" only applied when rule abstraction is not feasible (e.g., Juslin et al., 2003; Platzer & Bröder, 2013). However, given the similarity between tasks, it is quite plausible that exemplar-based processes may also affect numerical judgments in the presence of rules (e.g., von Helversen et al., 2014). Perhaps, the influence of exemplar-based processes in judgment has been underestimated because of the methods used for diagnosing exemplar-based processes: They either rely on an index measuring the ability to extrapolate or interpolate (e.g., Juslin et al., 2003) or on a classification of participants based on the best-fitting strategy (see Bröder et al., 2010; Juslin et al., 2003; Persson & Rieskamp, 2009; Platzer & Bröder, 2013). This coarse-grained analysis cannot detect subtle mixes of both processes if they exist.

1.2 RulEx-J — A simple mixture model for measurement

As characterized above, the diagnosis of individual strategies has largely relied on classifications that assign observed data patterns of individuals to the best-fitting of a set of strategies. The fit criterion may be some sort of scoring rule (e.g., Persson & Rieskamp, 2009) or the Maximum-Likelihood principle (e.g., Bröder & Schiffer, 2003). Both procedures, however, entail the simplistic assumptions that individuals use only one strategy across all trials and within trials. This simplistic analysis has both a theoretical and a pragmatic disadvantage: At the theoretical level, the demonstrations of ubiquitous similarity effects renders the adequacy of "pure" rules unlikely as an explanation of behavior. At the more pragmatic level, a categorical diagnosis of individuals as using either rules or exemplars wastes potential information about relative degrees of exemplar- and rule-based processing. Hence, besides giving a potentially wrong impression about the nature of processing, precision and statistical power to detect correlates of the respective processes is sacrificed.

In the following, we propose the simple measurement model *RulEx-J* to estimate the relative contribution of rule-based and exemplar-based processing in numerical judgments.[1] The goal of the model is not to spell out a detailed account of the cognitive processes involved in single judgments, but to provide a simple to use tool for assessing the relative impact of cue-abstraction and similarity-based ex-

[1]The possibilities for useful acronyms are restricted. We use the acronym "RulEx" for "Rule-Exemplar-Model", but to avoid confusion with the conceptually different "Rule+Execption Model" RULEX by Nosofsky, Palmeri and McKinley (1994), we added the "J" for "judgment" also to emphasize the different areas of application.

https://doi.org/10.1017/S1930297500006513 Published online by Cambridge University Press

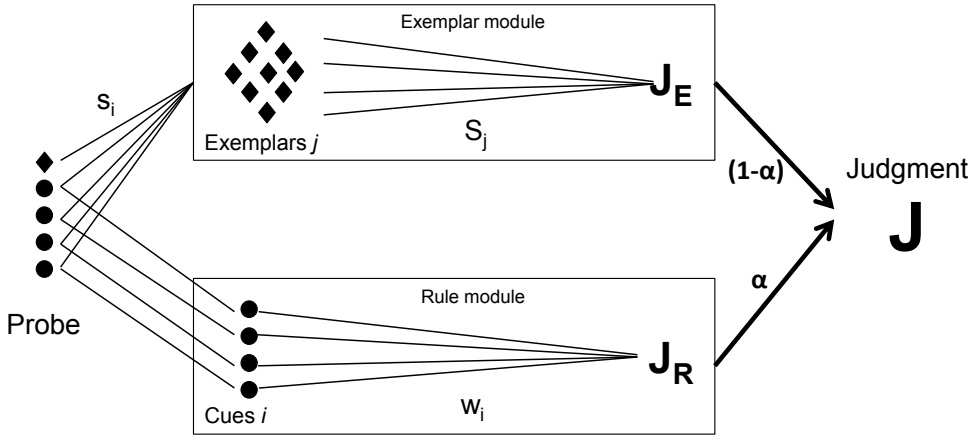

FIGURE 1: Schematic representation of the RulEx-J model. A to-be-judged probe is compared to exemplars stored in memory in the exemplar module (upper half), and the module creates a similarity-weighted judgment J_E from the stored criterion values. In the rule module (lower half), the probe is decomposed into cues that are integrated according to a weighted linear rule to generate the module's judgment J_R. Both tentative judgments are integrated into a final judgment J as a weighted average with relative weights α and $(1 - \alpha)$ given to the rule-based and exemplar-based judgments, respectively.

emplar retrieval on judgments. Hence, RulEx-J is intended as a pragmatic measurement model rather than an epistemic model (see Bröder, Kellen, Schütz & Rohrmeier, 2013, for a discussion of the distinction). However, each measurement model also has to specify a few theory-driven assumptions about the processes it intends to measure.

Figure 1 depicts a schematic outline of the model after some initial training has taken place. If a probe is presented to the judge (see left), it will be processed by two modules, the exemplar module E and the rule module R. The rule module will weigh the attributes/cues c_i of the probe with their respective weights w_i and produce an output J_R that is the weighted sum of cue values plus an additive constant. Note that this general rule resembles the linear integration of cues in Brunswik's (1955) lens model, and weights can be interpreted as cue utilizations. However, contrary to a common misunderstanding, the linear equation does *not* imply compensatory processing (see Bröder, 2000, for a discussion). Rather, it is a framework that can also model single-attribute-judgments without compensatory integration. In this specific case, all but one cue weights would get a weight of zero. Since the weights are estimated as free parameters, the linear equation can also represent unit weighting (Dawes & Corrigan, 1974) or noncompensatory weighting (Martignon & Hoffrage, 2002). Hence, the linear equation entails numerous cue abstraction rules as special cases, rendering the rule model very general. The core idea here is that the knowledge representation entails bivariate cue-criterion relations.

The exemplar model E contains all exemplars j encountered in the learning phase. The exemplar model is identical to Juslin and Persson's (2002) *ProbEx* model that is a simple extension of Medin and Schaffer's (1978) context model to numerical judgments. First, a similarity value S_j between the probe and each exemplar j is computed. The output J_E of the exemplar model with respect to the probe is then computed as the similarity-weighted average of criterion values of the available exemplars. The exact functions for obtaining the similarities and respective weighted averages are provided in the Appendix. An attention parameter s_i governs the impact of each cue dimension i on similarity, and for simplicity, the value is assumed to be constant across cues which has been a standard assumption in many judgment applications so far (Juslin & Persson, 2009; Juslin et al., 2003; Persson & Rieskamp, 2009; Platzer & Bröder, 2013).[2]

Hence, each module produces a tentative judgment as its output. Consider this as the module's best guess of a good judgment. Finally, both outputs are combined to produce the observed judgment J by a weighting parameter α that simply measures the relative impact of the rule versus the exemplar model:

$$J = \alpha \cdot J_R + (1 - \alpha) \cdot J_E \quad (1)$$

Hence, α is a mixture parameter that can be viewed as a measure of relative impact of both kinds of processes on the final judgments, yielding more fine-grained information than classifying a participant as using a strategy. Note, however, that purely rule- or exemplar-based processing are special

[2]In fact, many of the applications even assumed a constant $s = .5$. In the model presented here, the value of s is estimated from the data. With $s = 0$, the similarity is a step function, so only cue patterns identical to the probe are retrieved or considered. With $s = 1$, similarity is a flat function, meaning that similarity with regard to attribute values does not influence the impact of exemplars, and all exemplars are retrieved indiscriminately. Values between 0 and 1 model similarity-graded retrieval with smaller s implying a steeper dissimilarity function. The recovery simulations reported below showed that, for our data structure, the rule and exemplar function can readily be discriminated and parameters estimated if the maximum of s is set to .5.

https://doi.org/10.1017/S1930297500006513 Published online by Cambridge University Press

cases for $\alpha = 1$ and $\alpha = 0$, respectively. Thus, the advantage of the combined RulEx-J model compared to the separate modules increases as α approaches .5.

Hoffmann et al. (2013, 2014) used a conceptually very similar measure which they call a "strategy weight" for modeling final judgments. However, they did not yet apply the method to the learning phase and did not systematically validate the measure. The extension of their approach along these lines is the aim of the current article.

1.3 RulEx-J parameters and parameter estimation

For a judgment task with objects containing I attributes, RulEx-J has $(I + 3)$ parameters: I weights w_i (one for each cue i), one additive constant w_0 in the linear equation of the rule model, one attention parameter s for the exemplar model, and the mixture parameter α. Hence, a necessary condition for model identifiability is at least $I + 3$ judgments and sufficient variation in the cue patterns. For example, the parameters could be estimated from the last training blocks of a typical procedure in which participants repeatedly judge objects and receive feedback after each trial. Parameters can be estimated by minimizing the sum of squared deviations between predicted and observed judgments.

To estimate parameters for the RulEx-J model, the file "rulexj_functions.R" in the online supplement of the article includes a number of R functions for deriving model predictions and estimating parameters. Most importantly, the function *fit_rulexj* can be used to jointly estimate all RulEx-J parameters[3] for a number of participants that have provided judgments for a set of probes and have previously learned a set of exemplars. The optimization function underlying *fit_rulexj* strives to minimize the sum of squared deviations between the actual judgments and the predicted judgments using the optimization function *nlminb* in R (R Core Team, 2016). For each participant, it starts with a random set of parameter values and fitting is repeated a predetermined number of times with varying starting parameters (we used ten times for the following recovery simulations and 50 times for fitting the actual participant data), keeping the best fitting set of parameter values.

In addition, the *simulate_rulexj* function can be used to generate hypothetical judgments based on a randomly generated set of RulEx-J parameters. Using this function to simulate 1000 hypothetical participants each in a series of recovery simulations (without adding error to the generated judgments), we were able to recover the data generating parameters almost perfectly (see file "rulexj_recovery.html" in online supplement) using the following parameter restrictions: $.00001 \leq \alpha \leq .99999$, $0 \leq s \leq .50$, and $0 \leq w_i \leq 100$ for all weights in the rule module.[4] Besides, for s an upper bound of .5 was introduced as the exemplar module approaches constant predictions if s approaches 1. Note that the appropriate choice of the w_i boundaries depends on the range of values used in the judgment task. When simulating hypothetical participants in the recovery simulation, the sum of all w_i was additionally restricted to be ≤ 100, so that only judgments within the allowed range of the experiment would be produced. However, this limitation was not used when fitting the parameters. Still, in the later experiments the sum of w_i was ≤ 100 for all participants. These parameter restrictions were used when fitting parameters in all of the following experiments.

1.4 Overview of experiments

The first two experiments are validation studies manipulating strategies by instruction in a screen-based (Experiment 1A) or memory-based (Experiment 1B) environment. If α reflects the relative impact of rule-based reasoning, rule instructions are expected to yield higher average estimates of α than exemplar instructions. Such a construct validation of a measurement model must precede its application to substantive research questions. Experiment 2 explores whether adding pictures to cue profiles triggers exemplar processes that are, in turn, reflected in the α parameter.

All experiments followed the same standard procedure for investigating exemplar processes in judgment (see Pachur & Olsson, 2012): First, participants were trained in a *training phase* where they were repeatedly confronted with judgment objects (patients, bugs) characterized by cue patterns. These were judged on a relevant criterion (severity of illness, toxicity) and feedback about the actual criterion value was provided. In this phase, participants were expected either to extract a rule for prediction or to store the respective learned exemplars along with the criterion values. The last three out of the eight blocks of the learning phases are used to assess the model fit and to estimate the mixture parameter α. Note that both modules are nested within the more complex RulEx-J model, so the latter will always show a better model fit. Although we will also report Akaike weights (see Wagenmakers & Farrell, 2004) to correct for model complexity, this index only corrects for the number of parameters as a contributor to complexity, but not for the functional form. Hence, a more important question is whether the combined model improves the *predictions* of new behavior.

[3] The function additionally returns separate parameter estimates for the rule and exemplar modules. Besides, it offers the possibility to fit the α parameter after the other parameters.

[4] α was not allowed to reach its boundary values (0 and 1) as this led to problems in the estimation procedure due to the fact that at its boundaries changes in the parameters belonging to one of the modules did not have any effect on the fitting criterion (e.g., all rule parameters are irrelevant if $\alpha = 0$).

https://doi.org/10.1017/S1930297500006513 Published online by Cambridge University Press

TABLE 1: Structure of stimuli used in the experiments. The criterion values of the two cue patterns with deviations from the linear rule are set in boldface.

	Cue Pattern				**Pattern present in...**			**Criterion Value**
Pattern Number	Cue 1	Cue 2	Cue 3	Cue 4	Training	Decision	Test	
1	0	0	0	0			√	10
2	0	0	0	1	√		√	23
3	0	0	1	0	√		√	25
4	0	0	1	1		√	√	38
5	0	1	0	0	√	√	√	30
6	0	1	0	1	√	√	√	43
7	0	1	1	0		√	√	45
8	0	1	1	1		√	√	58
9	1	0	0	0	√	√	√	35
10	1	0	0	1		√	√	48
11	1	0	1	0		√	√	50
12	1	0	1	1	√	√	√	**70**
13	1	1	0	0		√	√	55
14	1	1	0	1	√		√	68
15	1	1	1	0	√		√	**63**
16	1	1	1	1			√	83

Therefore, two test phases without feedback followed: In the *first test phase* (henceforth: decision phase), participants made binary choice decisions between patterns. While some of these patterns were already present in the learning phase, other patterns were new. The *second test phase* again asked for numerical judgments, but now including new patterns that had not been trained in the learning phase. Of critical interest here are the most extreme cue patterns (1,1,1,1) and (0,0,0,0) that were never shown in the learning phase. Whereas a linear rule (with positive weights) predicts judgments that fall outside the learning range of criterion values, the similarity-weighted averaging of the exemplar model can never predict values outside the learning range. Hence, an observed inability to extrapolate is a marker of exemplar-based processes (see Juslin et al., 2003).

The criterion values associated with the cue patterns was a linear function of the form $Criterion = 25*Cue_1 + 20*Cue_2 + 15*Cue_3 + 13*Cue_4 + 10$ with the exception of patterns (1,0,1,1) and (1,1,1,0) for which the criterion values were switched. Hence, the environment had mostly a linear structure with two exceptions.[5] The logical structure of the stimuli as used in all experiments is depicted in Table 1.

[5]In a pilot study, we had used the same linear function with four exceptions which led to almost exclusive use of exemplar-based strategies and almost no rule abstraction.

2 Experiments 1A and 1B

The goal of Experiment 1 was the validation of the mixture parameter α by means of instruction. Cue patterns of judgment stimuli from two different fictitious content domains (patients or toxic bugs in Experiment 1A, only patients in Experiment 1B) were presented on the computer screen (Experiment 1A) or had to be retrieved from memory (Experiment 1B), and participants judged the hypothetical criterion, either with feedback (learning phase) or without (test phase). In the *rule* condition, participants were instructed before training to use feedback in order to learn the mathematical rule connecting cue and criterion values and to apply this rule to untrained objects later. In the *exemplar* condition, participants were told to memorize the objects and their criterion values during training and to judge untrained objects by their similarity to the memorized objects later. If participants at least partly followed this advice, a necessary requirement for the validity of RulEx-J would be a larger estimated α value in the rule condition as compared to the exemplar condition.

2.1 Method

Participants. Participants in Experiment 1A were 124 (96 female) students of the University of Mannheim who received a payment of €4 or fulfilled a course requirement.

https://doi.org/10.1017/S1930297500006513 Published online by Cambridge University Press

Sixty students of the University of Mannheim (44 female) participated in Experiment 1B for a compensation of €8.

Design. The main independent variable was the strategy instruction. This was manipulated between participants in both experiments. A second factor manipulated between participants only in Experiment 1A was the judgment material (patients vs. hypothetical bugs). This factor was used in order to demonstrate the validity of the model for different semantic embeddings of the task.

Materials. As a cover story, we used the established fictitious tropical disease task (Bröder et al., 2010; Gluck & Bower, 1988; Persson & Rieskamp, 2009) for half of the participants in Experiment 1A and all participants of Experiment 1B. In this task, participants had to judge the severity of a patient's disease based on a set of four binary symptoms (i.e., fever vs. hypothermia, hepatomegaly vs. cirrhosis, constipation vs. diarrhea, hypotension vs. hypertension). The other half of the participants in Experiment 1A were presented with fictitious bugs (Juslin et al., 2003) whose toxicity had to be judged based on four physical characteristics (i.e., long vs. short legs, gray vs. blue head, spotted vs. striped back, green vs. brown underside). Since experiment 1B was memory-based, the 16 patterns here were always presented together with a picture of a male person to facilitate memorizing.

Procedure. Participants were randomly assigned to one of the conditions. At the beginning of an experimental session participants received a consent paper that informed them that the study was concerned with the way people make decisions. Afterwards, the experimenter started the computer program.

In Experiment 1A, instructions informed participants that they will be presented with different patients (or bugs, respectively) whose severity of illness (or toxicity, respectively) has to be judged on a scale from 0 to 100. Depending on the condition, participants were either instructed to use the provided feedback about the correct values to learn a mathematical rule or to memorize the objects and their values. Following the instructions, participants were presented with the 64 training trials (eight training blocks each with eight objects). After the learning phase, the decision phase was introduced by instructing participants that in each of the following trials two old or new objects will be presented and that their task is to choose the object with the higher value (of illness or toxicity, respectively). Here, the previously learned rule (rule instruction) or the similarity to the memorized objects (exemplar instruction) should be used. The decision phase included ten patterns (four old, six new, see Table 1) and all pairwise combinations were presented resulting in 45 trials. Additionally, the six new patterns were paired a second time with each other (15 trials). Thus, the decision phase consisted of 60 trials in total. The final test phase covered 16 trials and instructed participants to judge the criterion values of old and new objects based on the learned rule or the similarity to the memorized objects, respectively. After the experiment was finished, participants were thanked, paid, and debriefed about the hypotheses underlying the present research. Sessions lasted about 20 minutes and included up to ten participants.

In Experiment 1B, the procedure was mostly the same as in Experiment 1A. However, since this experiment focused on *memory-based* decisions, there was a memorization phase before the learning phase. Here, participants were presented with 14 patients (the two extreme patterns were left out) and first had to guess the four symptoms of each pattern individually. After each trial, feedback about the correct cue values was provided. The program instructed participants to memorize the symptoms of all patients. If, after all 14 trials, people had memorized at least 45 out of the 56 information pieces (80%) correctly, they could continue with the next phase of the experiment. If participants failed to memorize enough information, they had to repeat the 14 trials. The presentation order was randomized for each new block. Because of the additional phase, sessions lasted about 50 minutes on average.

2.2 Results

Extrapolation and α values in the final test phase. In the final judgment phases of both experiments, participants judged all 16 cue patterns without feedback. From these 16 judgments, the α parameter can be estimated for each participant. In addition, the data allow for a qualitative check of exemplar-based processes if participants show a lack of extrapolation with respect to the most extreme cue patterns (0,0,0,0) and (1,1,1,1). If the instruction affected processing, extrapolation should be impaired in the exemplar conditions. Furthermore, if RulEx-J's α parameter is valid, it should yield lower values in the exemplar conditions.

Figure 2 shows the extrapolation patterns for Experiments 1A and 1B in the left and middle panels, respectively: Plot symbols represent participants' mean estimates of the criterion plotted against the actual criterion values. Whereas the most extreme patterns also yielded the most extreme judgments in the rule instruction conditions (filled circles), they were drawn towards the middle of the scale in the exemplar conditions (open circles) in both experiments. Hence, the result of this qualitative check is consistent with more exemplar-based processing in the exemplar instruction conditions than in the rule instruction conditions.

In terms of the model parameter α estimated from the final judgments, there was a corresponding difference with a lower mean value (α = .72, SD = .33) in the exemplar as compared to the rule condition (α = .81, SD = .26) in Experiment 1A. In a 2x2 between participants ANOVA with the factors *instruction* and *materials* all effects failed the conventional significance levels (instruction main effect: $F(1,120) = 3.41$, $p = .067$, $\eta_p^2 = .03$; materials main effect: $F(1,120) = 0.01$, $p = .909$, $\eta_p^2 < .01$; interaction: $F(1,120) = 3.86$, $p = .052$, η_p^2

https://doi.org/10.1017/S1930297500006513 Published online by Cambridge University Press

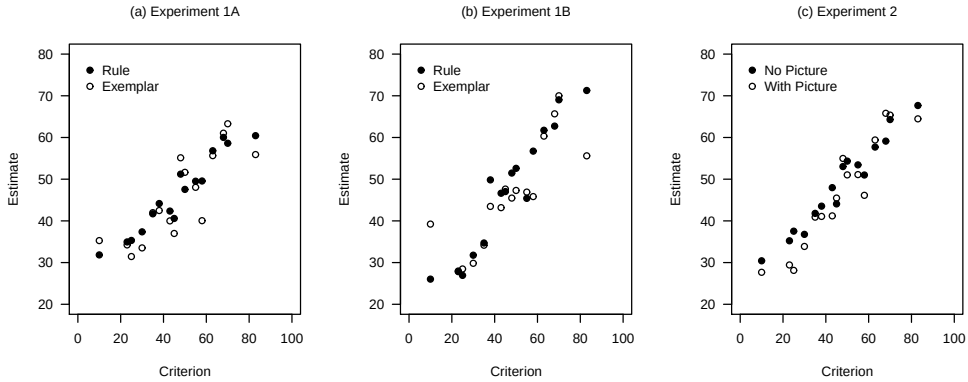

FIGURE 2: Mean final estimates in Phase 3 plotted against actual criterion values in (a) Experiment 1A, (b) Experiment 1B, and (c) Experiment 2. Exemplar conditions show no or less extrapolation for extreme criterion values of untrained patterns.

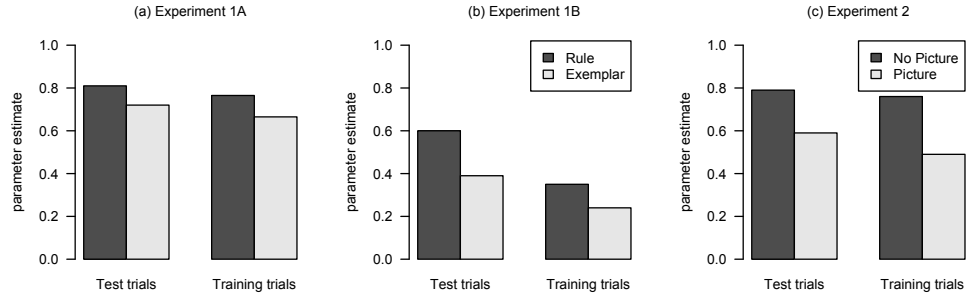

FIGURE 3: Mean estimated alpha values from all experiments. Estimates are either based on the final judgments including transfer stimuli (test trials) or the last training blocks (training trials). Dark bars denote experimental conditions favoring rules, light bars show conditions favoring exemplar use.

= .03). A further examination of the interaction showed that a significant main effect of instruction was observed with the patients materials (α_{rule} = .87, *SD* = .24, α_{exemplar} = .67, *SD* = .35, $t(53.14) = 2.66$, $p = .010$, $d = 0.68$), whereas there was no difference for the bugs (α_{rule} = .76, *SD* = .26, α_{exemplar} = .76, *SD* = .32, $t(58.15) = 0.08$, $p = .933$, $d = 0.02$). In Experiment 1B, the α value in the rule condition (α = .60, *SD* = .30) was also significantly higher than the one in the exemplar condition (α = .39, *SD* = .23), $t(53.70) = 3.01$, p = .004, $d = 0.78$.

Hence, the instruction had the intended effect to induce different strategies as reflected in the qualitative extrapolation. Importantly, this difference was adequately reflected in the model parameter α. Figure 3 shows the mean α estimates of Experiments 1A and 1B in the left and middle panels, respectively.

Correct decisions in the paired comparisons. As in the final judgments, rule extractors should be better able to generalize their knowledge to new patterns than exemplar-based decision makers also in the paired comparisons (see Pachur & Olsson, 2012). Figure 4 shows the percentage of correct decisions as a function of the number of new patterns in a trial and the experimental condition in Panels (a) and (b) for both experiments, respectively.

Apparently, participants were better when two old patterns had to be compared if they had received an exemplar as opposed to a rule instruction. However, the ranking switched when two new patterns had to be compared. A 2x3 mixed ANOVA revealed a main effect of pattern novelty, $F(2,121)$ = 23.86, $p < .001$, η_p^2 = .28, and a significant interaction of novelty and instruction, $F(2,121) = 5.72$, p = .004, η_p^2 = .09, in Experiment 1A. The main effect of instruction did not reach significance, $F(1,122) = 0.97$, p = .328, η_p^2 = .01. Hence, the paired comparisons mirror the judgment extrapolation results and replicate the pattern reported by Pachur and Olsson (2012). The 2x3 mixed ANOVA for Experiment 1B also showed a significant main effect of pattern novelty, $F(2,57) = 144.11$, $p < .001$, η_p^2 = .84, but no significant main effect of instruction, $F(1,58) = 0.41$, p = .523, η_p^2 = .01. The interaction of novelty and instruction failed to reach the conventional significance criterion, $F(2,57) = 2.90$, p = .063, η_p^2 = .09. In sum, however, results are generally as expected with better generalization ability to new objects if participants were instructed to learn a rule rather than if they memorized specific exemplars.

Modeling the final training blocks. Tables 2 and 3 report the residual sums of squares for the rule model, the exemplar model, and the combined RulEx-J model in both experi-

https://doi.org/10.1017/S1930297500006513 Published online by Cambridge University Press

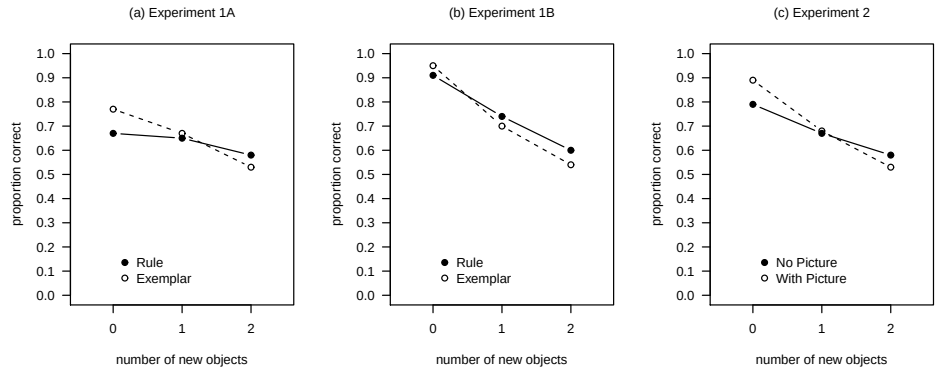

FIGURE 4: Percentage of correct choices in decision phases of (a) Experiment 1A, (b) Experiment 1B and (c) Experiment 2, plotted as a function of the number of new patterns in a pair and the experimental conditions.

TABLE 2: Mean residual sum of squares and mean Akaike weights for the three models in Experiment 1A separately for each instruction condition (rule vs. exemplar) and across both conditions (overall).

	Residual Sum of Squares			**Mean Akaike Weights**		
	Rule	Exemplar	Overall	Rule	Exemplar	Overall
Rule model	3770	2688	3229	.44	.42	.43
ProbEx model	5474	4154	4814	**.47**	**.50**	**.48**
RulEx-J model	**3719**	**2609**	**3164**	.09	.08	.09

Note. The best fitting model according to the respective fit criterion is highlighted in bold. In five cases in the exemplar condition one or more residual sum of squares had a value of zero making the Akaike weight undefined. Therefore, these cases were excluded for the calculation of the mean Akaike weights.

ments, respectively. In both conditions of both experiments, more variance was explained by the Rule model than by the ProbEx model. Since both models are nested in RulEx-J, it is unsurprising that the latter yields the best fit in all conditions. The tables also list mean Akaike weights for the three models to correct for the higher complexity as indicated by the number of parameters in each model. In terms of Akaike weights, the exemplar model outperformed the others in all conditions of Experiment 1A as well as in Experiment 1B.

Alpha values based on the training phase. We also estimated α from the last three training blocks (24 trials with eight patterns altogether). For Experiment 1A, a 2x2 ANOVA yielded the expected main effect of instruction, $F(1,120) = 3.94$, $p = .049$, $\eta_p^2 = .03$, and no main effect of materials, $F(1,120) = 0.10$, $p = .748$, $\eta_p^2 < .01$. However, there was a significant interaction of both factors, $F(1,120) = 6.41$, $p = .013$, $\eta_p^2 = .05$. A further examination showed that a main effect of instruction was observed only with the patients materials ($\alpha_{\text{rule}} = .87$, $SD = .25$, $\alpha_{\text{exemplar}} = .60$, $SD = .38$, $t(51.52) = 3.34$, $p = .002$, $d = 0.85$), whereas there was no significant difference with the bugs ($\alpha_{\text{rule}} = .70$, $SD = .32$, $\alpha_{\text{exemplar}} = .73$, $SD = .37$, $t(58.73) = 0.37$, $p = .712$, $d = 0.09$). Hence, the validation of α from the final trials of the learning phase was successful overall, but a closer inspection showed that this was restricted to the patient materials. In Experiment 1B, the α value based on the last training blocks was also descriptively, but not significantly higher in the rule ($\alpha = .35$, $SD = .30$) than in the exemplar instruction condition ($\alpha = .24$, $SD = .36$), $t(58) = 1.30$, $p = .198$, $d = 0.34$. Also, both α values were obviously lower than in experiment 1A indicating the expected tendency towards more exemplar-based processing in memory-based decisions and judgments.

Model predictions. RulEx-J yields a better fit in terms of residuals than both the rule and the ProbEx model because it is the more general model, but it was not superior in terms of Akaike weights. Does the additional parameter unduly increase model flexibility due to overfitting? The ultimate test here is to predict new behavior with the estimated parameters. The model parameters estimated from the last three training blocks of each individual were used to predict individual behavior in both test phases: decisions and final judgments. For the decisions, criterion values of both options were predicted with the individual's parameter values

https://doi.org/10.1017/S1930297500006513 Published online by Cambridge University Press

TABLE 3: Mean residual sum of squares and mean Akaike weights for the three models in Experiment 1B separately for each instruction condition (rule vs. exemplar) and across both conditions (overall).

	Residual Sum of Squares			**Mean Akaike Weights**		
	Rule	Exemplar	Overall	Rule	Exemplar	Overall
Rule model	1679	1035	1357	.20	.10	.16
ProbEx model	2098	2240	2169	**.57**	**.76**	**.65**
RulEx-J model	**1528**	**893**	**1210**	.23	.14	.19

Note. The best fitting model according to the respective fit criterion is highlighted in bold. In 14 cases one or more residual sum of squares had a value of zero making the Akaike weight undefined. Therefore, these cases were excluded for the calculation of the mean Akaike weights.

TABLE 4: Correctness of model predictions for decisions and judgments in Experiment 1A separately for each instruction condition (rule vs. exemplar) and across both conditions (overall).

	Mean Consistency Between Model Predictions and Actual Decisions			**Mean Difference Between Model Predictions and Actual Judgments**		
	Rule	Exemplar	Overall	Rule	Exemplar	Overall
Rule model	**.694**	**.684**	**.689**	11.14	11.23	11.18
ProbEx model	.622	.665	.643	11.91	11.67	11.79
RulEx-J model	.680	.672	.676	**10.95**	**10.54**	**10.74**

(from fitting the Rule model, ProbEx, and RulEx-J, respectively), and the object with the higher value was predicted to be chosen. This was compared to the empirical choice, and the percentage of matches was assessed for each participant and (sub-)model. Similarly, the final judgments of a participant at the end of the experiment were predicted with the parameters estimated from the training phase, and mean absolute deviations were assessed. Tables 4 and 5 show the mean accuracy of the three model predictions for both the decisions (percent correct) and the final judgments (mean absolute deviation).

Table 4 shows that RulEx-J was significantly the best model to predict final estimates in all conditions of Experiment 1A, ts > 2.31, ps ≤ .024, whereas choices were better (albeit not significantly) predicted by the rule model. In Experiment 1B (see Table 5), results were not as clear-cut: Here, RulEx-J outperformed the other models in predicting the final estimates (including the new patterns) in the rule condition, whereas the exemplar model led to better predictions in the exemplar condition. Choices were better predicted by the rule model in the rule condition and by the exemplar model in the exemplar condition. Except for this latter observation, the results suggest that the inclusion of the additional parameter in RulEx-J does *not* lead to an overly flexible model that overfits unsystematic aspects of the data. Note that in Experiments 1A and 1B, choice predictions of RulEx-J were on average less accurate than the better submodel by a margin of only .013 and .003, respectively. In case of severe overfitting, one would have expected a more severe loss in prediction accuracy.

Stability of strategy. Finally, the α estimates from the final learning blocks and from the final judgments were correlated, yielding significant correlations, $r = .517$, $p < .001$ and $r = .300$, $p = .020$, in both experiments, respectively. As a measure of behavioral consistency, we correlated each participant's judgments for the eight repeated patterns from the last training block and from the final judgment phase, yielding fairly high median correlations of .70 and .99 in Experiments 1A and 1B, respectively (means: .57 and .89).

2.3 Discussion

Experiments 1A and 1B aimed at validating the RulEx-J model by manipulating training instructions. In line with expectations, this leads to different strategies on the aggregate level as documented qualitatively by the extrapolation results. More importantly, this difference was adequately reflected in the mixture parameter α, regardless whether its value was estimated from the last training blocks or the final judgments. However, it is at present unclear why the expected effect was absent for the "bug" materials in Experiment 1A.

When predicting final judgments with the model parameters estimated from the learning trials, the RulEx-J model

https://doi.org/10.1017/S1930297500006513 Published online by Cambridge University Press

TABLE 5: Correctness of model predictions for decisions and judgments in Experiment 1B separately for each instruction condition (rule vs. exemplar) and across both conditions (overall).

	Mean Consistency Between Model Predictions and Actual Decisions			**Mean Difference Between Model Predictions and Actual Judgments**		
	Rule	Exemplar	Overall	Rule	Exemplar	Overall
Rule model	**.653**	.633	.643	8.19	10.30	9.24
ProbEx model	.631	**.660**	**.645**	9.09	**8.00**	8.55
RulEx-J model	.642	.642	.642	**8.07**	8.30	**8.19**

outperformed the models that assume only one process for final judgments (with the exception of the exemplar condition in Experiment 1B). However, choices were always (albeit not significantly) better predicted by one of the component models. The overall predictive success concerning the choice was not overwhelming (65-70%). However, if the better fit of RulEx-J as compared to the component models was due to overfitting, one would have expected *impaired* predictions for the choices and final estimates which was clearly not the case.

Finally, the α parameter estimates were lower in the second experiment (particularly when estimated from the learning phase) which would be expected if a memory-based procedure is paired with the specific cue format used here ("alternative" cues, see Bröder et al., 2010; Platzer & Bröder, 2013). Also, the generally high α values in the first experiment reflect people's preference of rule-based strategies when cue abstraction is possible (Bröder et al., 2010; Hoffmann et al., 2013; Juslin et al., 2003; Karlsson et al., 2008; Platzer & Bröder, 2013). Although comparisons across experiments have to be interpreted cautiously, this is a further hint to the model's validity. Alternatively, the difference between the two experiments could have been caused by the addition of pictures in experiment 1B which could have triggered exemplar-based processing. This hypothesis was further tested in Experiment 2.

3 Experiment 2

Although the reported experiments were an important first step to validate the measurement method, they relied on instructed strategy use. This may have induced demand effects or other unknown differences to situations with spontaneous strategy selection, perhaps artificially influencing the α values. Hence, the aim of Experiment 2 was to apply the model to a substantive research question: Does the inclusion of pictures encourage exemplar-based processing? We hypothesized that presenting cue patterns with individuating pictures (here: portraits) would encourage exemplar storage as opposed to mere descriptions of the cue patterns. If this was the case, we should observe the respective change in the α parameter.

Second, one problem with the current implementation of the model is that the parameter estimates from the learning phase may overestimate exemplar use. The reason is that the model cannot distinguish between persons who learned all exemplars and participants who learned the rule plus the two exceptions from the rule. Both would eventually show perfect performance in the last blocks of training where all objects had been encountered and participants had received feedback on each object. A hint to this potential problem is that α estimates from the learning phase were lower than estimates from the final test phase (with transfer stimuli) in Experiment 1B.

Hence, in the current experiment, we compared a picture with a no picture condition in a screen-based environment. Also, we included the two extreme patterns (0,0,0,0) and (1,1,1,1) in the learning phase, but without feedback about the criterion values. This was done to better discriminate between rule learners and exemplar users in the final blocks of the training phase.

3.1 Method

Participants. Participants were 60 (42 female, 18 male) students of the University of Mannheim who received a payment of €4 or fulfilled a course requirement.

Design. Presentation format (without pictures vs. with pictures) was the only independent variable varied between participants. Analyses were based on the training with the extreme patterns included.

Materials. The fictitious tropical disease task was used for all participants. In the picture condition, the 16 patterns were always accompanied by a picture of a male person.

Procedure. The procedure was mostly the same as in Experiment 1A. Since the two extreme patterns were included in the learning phase, this phase now consisted of 80 instead of 64 trials (eight training blocks each with ten objects). However, in contrast to the other patterns, no feedback was given about the criterion values of the extreme patterns. Sessions lasted about 25 minutes and included up to ten participants in a room with computer cubicles.

https://doi.org/10.1017/S1930297500006513 Published online by Cambridge University Press

3.2 Results

Extrapolation and α values in the test phase. The α value in the no picture condition (α = .79, *SD* = .31) was significantly higher than the α value in the condition with pictures (α = .59, *SD* = .31), $t(58) = 2.50$, $p = .015$, $d = 0.65$ (see Figure 3, right panel). Participants in both conditions were able to extrapolate. However, this ability was stronger in the condition without pictures (see Figure 2, right panel).

Correct decisions in the paired comparisons. Decision trials with two old patterns as well as trials with one old pattern were more correctly solved by participants in the picture condition. However, participants in the condition without pictures were more accurate in trials with two new patterns. The 2x3 mixed ANOVA revealed a significant main effect of pattern novelty, $F(2,57) = 58.73$, $p < .001$, η_p^2 = .67, and a significant interaction of novelty and presentation format, $F(2,57) = 3.46$, $p = .038$, $\eta_p^2 = .11$. The main effect of instruction did not reach significance, $F(1,58) = 0.96$, p = .332, $\eta_p^2 = .02$. Figure 4 (right panel) shows the percentage of correct decisions for the paired comparisons.

Modeling the final training blocks. Regarding the residual sum of squares, the RulEx-J model achieved the best fit in both conditions. The Rule model explained more variance than the ProbEx model overall and in the no picture condition, whereas the Rule model and the ProbEx model explained nearly the same amount of variance in the condition with pictures. The mean Akaike weights revealed that the ProbEx model had the best fit in the picture condition, whereas the Rule and the ProbEx model had an approximately equal fit in the condition without pictures. Table 6 reports the residual sum of squares as well as the mean Akaike weights for the three models based on the learning phase with the extreme patterns included.

Alpha values based on the learning phase. The α value based on the last training blocks was also significantly higher in the no picture (α = .76, *SD* = .30) than in the picture condition (α = .49, *SD* = .36), $t(55.71) = 3.07$, $p = .003$, d = 0.79. Without the inclusion of the extreme patterns, the α value still was significantly higher in the no picture (α = .74, *SD* = .37) than in the picture condition (α = .50, *SD* = .42), $t(56.90) = 2.27$, $p = .027$, $d = 0.59$.

Model predictions. Descriptive data for the decisions showed that RulEx-J made the best predictions in both conditions. However, the model factor failed to reach conventional significance in a 2x3 (presentation format x model) mixed ANOVA, $F(2,55) = 2.67$, $p = .078$, $\eta_p^2 = .09$. Neither the main effect of the presentation format, $F(1,56) = 0.14$, p = .711, $\eta_p^2 < .01$, nor the interaction was significant, $F(2,55)$ = 0.01, $p = .990$, $\eta_p^2 < .01$. Regarding the final judgments, RulEx-J made the best predictions overall and in the no picture condition, but was slightly outperformed by the ProbEx model in the picture condition. The 2x3 ANOVA proved the main effect of the model factor to be significant, $F(2,57)$ = 23.74, $p < .001$, $\eta_p^2 = .45$. The main effect of the presentation format, $F(1,58) = 1.04$, $p = .312$, $\eta_p^2 = .02$, and the interaction were not significant, $F(2,57) = 1.95$, $p = .151$, η_p^2 = .06. A closer look at the main effect of the model factor showed that there was no significant difference between the RulEx-J and the ProbEx prediction accuracy, $t(59) = 0.49$, $p = .627$, $d = 0.06$, while the other two paired comparisons (Rule vs. ProbEx and Rule vs. RulEx-J) were significant, *p*s < .050. Without the extreme patterns in the learning phase, the results for decisions were similar, whereas the results for judgments were more in favor of RulEx-J: Here, RulEx-J made the best judgment predictions in both conditions. Table 7 shows the mean accuracy of the three model predictions for both the decisions and the final judgments based on the learning phase with the extreme patterns included.

Stability of strategy. The α estimates from the final learning blocks without the extreme patterns and from the final judgments were significantly correlated, $r = .644$, $p < .001$. The correlation between the α values from the final learning blocks with the extreme patterns and from the final judgments was also significant, $r = .705$, $p < .001$. The median correlation between judgments for the ten repeated patterns from last training block and final judgment phase was .70 (mean: .66).

3.3 Discussion

Experiment 2 investigated the influence of pictures on the relative impact of rule-based learning and explored an alternative way of calculating the α value by including patterns without feedback in the learning phase. As expected, the α parameter and thus the extent of rule-based processing was higher when no pictures were presented together with the objects. This effect was also reflected in the decision accuracy and the extrapolation ability and is consistent with the results of the previous experiments. Pictures thus indeed seem to trigger more exemplar-based processing. However, it is premature to conclude that this effect holds for all pictures or just for portraits of people, since the latter was the only kind of pictures we used in our experiments. We speculate that pictures may foster exemplar-based processing if they allow for "individuating" cue patterns as repeated instances.

Including transfer stimuli without feedback already in the training phase improved the prediction accuracy of RulEx-J: In Experiment 2, RulEx-J almost consistently outperformed both the rule model and the exemplar model alone in predicting the choices as well as the final estimates. This is also mirrored in a higher correlation of the α values as estimated from training and test.

RulEx-J was the best model for predicting both decisions and judgments, although the ProbEx model performed similarly well for the judgments. These results did not substantially change when the analyses were repeated without the extreme patterns in the learning phase. This substanti-

https://doi.org/10.1017/S1930297500006513 Published online by Cambridge University Press

TABLE 6: Mean residual sum of squares and mean Akaike weights for the three models for each presentation format condition (without vs. with pictures) and across both conditions (overall).

	Residual Sum of Squares			**Mean Akaike Weights**		
	Without Pictures	With Pictures	Overall	Without Pictures	With Pictures	Overall
Rule model	5109	4250	4680	**.46**	.15	.31
ProbEx model	6725	4251	5488	.45	**.76**	**.61**
RulEx-J model	**4786**	**3424**	**4105**	.08	.09	.09

TABLE 7: Correctness of model predictions for decisions and judgments separately for each presentation format condition (without vs. with pictures) and across both conditions (overall).

	Mean Consistency Between Model Predictions and Actual Decisions			**Mean Difference Between Model Predictions and Actual Judgments**		
	Without Pictures	With Pictures	Overall	Without Pictures	With Pictures	Overall
Rule model	.675	.686	.680	11.21	10.89	11.05
ProbEx model	.668	.670	.669	10.81	**9.70**	10.25
RulEx-J model	**.685**	**.687**	**.686**	**10.59**	9.73	**10.16**

ates the advantage of the RulEx-J model over the other two models. It is important to note, however, that this predictive advantage (generalization to new data) was *not* adequately captured in the Akaike weights which only correct for the number of parameters but do not use an external criterion for predictive accuracy.

4 Joint analyses

In the introduction, we stated that, in addition to the theoretical advantage of the model, the continuous α parameter characterizing a person may contain more information than a coarse classification to the "best" strategy. Analyzing all experiments together to increase statistical power, we see an overall difference in α between the "rule" and "exemplar" conditions (.67 vs. .52, respectively, $t(242) = 3.15$, $p = .002$) confirming the result from the separate experiments. However, classifying participants to the rule-based or exemplar-based strategy due to the lowest AIC[6] from the last training blocks yields the same result, showing more apparent "exemplar" users in the exemplar conditions (75.4%) than in the rule conditions (54.9%, $\chi^2(1) = 11.28$, $p = .001$). That is, qualitative results are similar with both methods. Hence, one may ask if there is any additional predictive power of using α instead of a classification. Although the experiments were not initially designed to test this matter, we explored in two analyses whether (a) the α parameter predicts additional aspects of behavior such as the degree of extrapolation in the final judgments and (b) whether it helps to predict the experimental condition of a participant. We used all experiments simultaneously to increase the power of the analysis with experiment coded as two dummy variables.

First, we analyzed the extrapolation behavior in the final estimates, using each participant's range of judgments (maximum-minimum) as a proxy measure for his or her degree of extrapolation.[7] A regression analysis involving experiment (two dummy predictors) and strategy classification based on the lowest AIC (exemplar vs. rule) yielded $R = .27$ ($p < .01$) which increased to $R = .30$ when α estimates were included as additional predictors, the difference in R being significant, $F(1,239) = 4.71$, $p = .03$. Hence, α as estimated from the training phase yields some predictive power for the final judgments in addition to a mere strategy classification.

Second, we ran a binary logistic regression to predict the experimental condition (fostering either exemplar- or rule use) of each participant in the three experiments. In the first block of the regression, two dummy variables coding the three experiments and the range of final judgments as a proxy for extrapolation were entered, followed by the strategy classification in the second block. In the final block, α was entered as a predictor to test whether it yields additional prediction accuracy. In the final model, both range and α

[6]For 19 participants in Experiments 1A and 1B, AIC was undefined due to a perfect fit of the Exemplar model and $\alpha = 0$. These participants were classified as exemplar users.

[7]Results are the same if other indices of extrapolation are used, e.g., the amount of exceeding the actual criterion values of the extreme patterns or the number of times extreme values were exceeded by the judgments.

https://doi.org/10.1017/S1930297500006513 Published online by Cambridge University Press

yielded significant regression weights with Wald statistics of 3.84 (p = .050) and 4.49 (p = .034), respectively, whereas the strategy classification fell short of a conventionally significant influence (Wald statistic of 3.61, p = .058).

Hence, these analyses yield first evidence that judgment analysis with RulEx-J may predict some variance in behavioral data in addition to a mere strategy classification.

5 General Discussion

In this paper, we proposed a simple and easy-to-use method to measure the relative impact of rule-based and exemplar-based processes in judgment. In an extension to similar methods introduced by Hoffmann et al. (2013, 2014), we provided a construct validation of the mixing parameter. The RulEx-J model's mixture parameter α responded adequately to strategy instructions (Experiments 1A and 1B) and to the inclusion of individuating pictures accompanying the cue patterns (Experiment 2). Although the model fit of RulEx-J in terms of Akaike weights was not consistently superior to the component modules, the prediction of choices and final estimates (including transfer stimuli) was almost equal or better. The sensible behavior of α together with these prediction results confirms that the inclusion of this parameter captures systematic aspects of the data rather than unsystematic noise due to overfitting. Predicting later behavior (choices and judgments for new exemplars) by the parameters estimated from the learning trials is in our view the most convincing test for showing that the additional parameter does not induce merely noise-fitting flexibility. Here, the more complex model fared at least as well as the component models. Hence, compared to this cross-prediction results, the penalty of Akaike weights for the extra parameter seems too strict in this case.[8] In addition to providing a tool for measurement, this pattern also seems to validate the idea of a process mix in cue-based judgments in line with former research (e.g., Hahn et al., 2010; von Helversen et al., 2014).

5.1 Interpretation of model parameter α

In the introduction, we made the processing assumption that both the ProbEx and the Rule submodel run in parallel, and the respective results are integrated in each judgment as a weighted average. This idea is in line with models proposed by Herzog and von Helversen (in press) or Juslin et al. (2008). For example, Herzog and von Helversen (in press) emphasize that "blending" two different processes with differing strengths and weaknesses might be ecologically rational analogous to the "wisdom of the crowd" phenomenon in which averaging across independent judges increases judgment accuracy (e.g., Herzog & Hertwig, 2014).

However, an open question is if a trialwise switching between strategies (rather than averaging the results of the two models in each trial) could be mimicked by the RulEx-J model as well? The answer is both "yes" and "no". In additional recovery simulations (see file "rulexj_recovery_add_sim.html" in the supplement), we simulated participants that repeatedly provided judgments for the exemplars from the training phase (with randomly drawn parameters using the same parameter restrictions as in the initial recovery simulation). In one simulation, for each participant 50% of the judgments were perfectly in line with the ProbEx submodel and 50% perfectly in line with the Rule submodel. In a second simulation, 75% of the judgments followed the ProbEx submodel and 25% the Rule submodel, and in a third simulation 25% of the judgments followed the ProbEx submodel and 75% the Rule submodel. Next, the RulEx-J model was fitted to these perfectly mixed judgments. In the vast majority of cases, the estimated α precisely corresponded to the relative frequency of ProbEx and Rule judgments, with a value of .50 in the first, a value of .25 in the second, and a value of .75 in the third simulation. Besides, the data generating parameters of the submodels were also recovered extremely well. Hence, the α parameter may also mimic trialwise switching and estimate the proportion of trials in which one or the other strategy was used.

However, although the *parameter* mimics strategy switches, the *model fit* does not. In our simulation, the sums of squared deviations were extraordinarily high when analyzing the switching data with the combined model. Since there are different benchmarks for good fits dependent on the stimulus structure, we cannot give general recommendations how to assess when a fit is "bad enough" so that the judgments could also stem from trial mixing rather than process mixing. In many contexts, one will be interested in more general statements about rule-based versus exemplar-based processing with less emphasis on the specifics of how strategies are mixed (within or across trials). If one is interested in characterizing the mixing more exactly, however, one may use bootstrap sampling to generate trial mixes according to the model parameters estimated for a person and by comparing the empirical model fit to the fit distribution generated by data assuming across-trial mixing.

[8] Although assessing model flexibility by using formal methods like Minimum Description Length (MDL) or Normalized Maximum Likelihood (NML) was beyond the scope of this paper, an informal assessment of model flexibility was conducted by simulating visualizations of (2D-projections of) the prediction spaces of the component models as well as the joint model RulEx-J (with α = .50) which show that the prediction space of RulEx-J does not extend beyond the boundaries of the rule component. The simulations are documented in "rulexj_prediction_space_2d_plots.html" in the online supplement.

https://doi.org/10.1017/S1930297500006513 Published online by Cambridge University Press

5.2 Experimental prerequisites for using the model

One prerequisite for using RulEx-J is that experimental stimuli and criterion values are chosen in a way that results in partially non-overlapping prediction spaces of the submodels. If both submodels' prediction spaces were identical, α would not be identifiable. However, diagnostic data situations with differential predictions of strategies are of course also necessary if a classification of participants to "strategies" is intended (Bröder & Schiffer, 2003; Persson & Rieskamp, 2009). How much should the predictions differ in order to estimate α accurately? We cannot suggest a quantitative metric to define this difference between predictions for our model, but a general suggestion is that greater differences allow for more diagnostic data to disentangle the processes. In our experiments, the criterion values of the 16 patterns could be predicted almost perfectly by a linear rule with the exception of two "switched" values that violated the rule. Despite this small difference, the model behaved well. Hence, we are optimistic that useful parameter differences can be found even with subtle differences in model predictions. Note that the use of extreme differences to maximize diagnosticity may also backfire if environments are created that can only be tackled by one process, thus leading to floor or ceiling effects in α.[9]

5.3 Limitations

There are limitations of the method. First, the model used here is certainly simplistic and probably does not describe the actual cognitive processes that lead to a judgment. This is especially true for the rule model which sticks to a "paramorphic" representation of the judgment process (Hoffman, 1960). Also, the model assumes a constant mixture of processes in each judgment, which is probably unrealistic (Erickson & Kruschke, 1998). However, our intention is not to propose RulEx-J as an *epistemic* model in this sense, but rather as a *pragmatic* measurement tool. Pragmatic models draw on theoretically informed principles and thus provide information superior to "theory-free" metrics, but they are not intended to model the process in detail. Rather, they may provide more sophisticated or sensitive measures of processing than surface statistics do (see Bröder et al., 2013, for a discussion). In this vein, they can serve as tools for hypothesis testing (e.g., whether some variable influences the processing mode in general) and thus indirectly inform the development of truly epistemic models. With the empirical demonstrations in this article, we showed that RulEx-J may be an informative and easy-to-use tool in the sense of a pragmatic model.

[9]In pilot work, we used a condition with four switched criterion values in eight training patterns. This led to almost exclusive use of exemplar processes in all participants which of course counteracted our goal to manipulate the α parameter between conditions.

A second limitation of the model is the interpretation of the mixture parameter. It is tempting to interpret α as a "proportion" of rule-based relative to exemplar-based processes going on. However, a closer look reveals that this view is not tenable from a philosophical point of view (unless you specify the unit in which you measure the "amount" of a process). More trivially, however, the estimate of α will depend on the actual set of stimuli used for estimation since different sets of cue patterns and criterion values will differ in their ability to differentiate between the kinds of processes involved. Hence, we caution researchers to interpret the absolute values of α to declare a preponderance of this or that process. What is unproblematic, however, is the comparison of α across experimental conditions using stimuli of similar logical structure to assess the impact of experimental manipulations on the mode of processing analogous to our experiments reported here.

Third, the model as presented here is currently a static model that does not include learning processes. It may be used to *describe* the learning process by applying the model to adjacent training blocks, but it currently lacks a description of the learning mechanism itself. This is a desideratum for a future development of a more epistemic model.

Fourth, one potential bias arises if participants fail to learn systematic behavior in the training phase and respond randomly. In this case, the parameter estimate of α will tend to be biased towards 1.0 since the 5-parameter rule module is more flexible to overfit noise than the 1-parameter exemplar module (see file "rulexj_recovery_add_sim.html" in the supplement). Hence, we recommend eliminating participants with a zero correlation or less between judgments and actual criterion values at the end of the training phase since this would certainly indicate that nothing was learned.

Similarly, the median correlations between repeated judgments for the same objects show that there was substantial consistency in behavior, but still, the reliability was less than perfect (median correlations of .70, .99, and .70) with considerable interindividual variation. In fact, some people showed very low or even negative correlations, indicating a lack of any systematic strategy (maybe due to missing motivation). The model's prediction success is certainly limited to systematic behavior. Therefore, the reliability achieved in a person's repeated judgments defines an upper limit of the model's ability to describe the data.

5.4 Summary and conclusion

To summarize, we conjecture that the more fine-grained analysis allowed by the mixture parameter in comparison to strategy classifications may enable the detection of subtler influences on the processing mode. In the online supplementary materials, we provide all data and the R code for parameter estimation as well as the parameter recovery simulations. We encourage researchers to re-analyze their existing data sets with the method if they fit the data structure needed.

https://doi.org/10.1017/S1930297500006513 Published online by Cambridge University Press

References

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

Bröder, A. (2000). A methodological comment on behavioral decision research. *Psychologische Beiträge, 42*(4), 645–662.

Bröder, A., Kellen, D., Schütz, J., & Rohrmeier, C. (2013). Validating a two-high-threshold measurement model for confidence rating data in recognition. *Memory, 21*, 916–944.

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*, 326-338.

Bröder, A., & Schiffer, S. (2003). "Take The Best" versus simultaneous feature matching: Probabilistic inferences from memory and effects of representation format. *Journal of Experimental Psychology: General, 132*, 277–293.

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

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

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95–106.

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

Gigerenzer, G., & Goldstein, D. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*, 650–669.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117*, 227–247.

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

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

Herzog, S. M., & Hertwig, R. (2014). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences, 18*(10), 504–506.

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

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin, 57*, 116–131.

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

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

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

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

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

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

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22–44.

Kruschke, J. K., & Erickson, M. A. (1994). Learning of rules that have high-frequency exceptions: New empirical data and a hybrid connectionist model. In A. Ram & K. Eiselt (Eds.), *Proceedings of the sixteenth annual conference of the cognitive science society* (pp. 514–519). Hillsdale, NJ: Erlbaum.

Martignon, L., & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision, 52*, 29–71.

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

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

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

Persson, M., & Rieskamp, J. (2009). Inferences from memory: Strategy- and exemplar-based models compared. *Acta Psychologica, 130*, 25–37.

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

R Core Team. (2016). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

https://doi.org/10.1017/S1930297500006513 Published online by Cambridge University Press

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

Thibaut, J.-P., & Gelaes, S. (2006). Exemplar effects in the context of a categorization rule: Featural and holistic influences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 1403–1415.

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

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11*, 192–196.

Appendix: The exemplar model

The exemplar model in RulEx-J is equivalent to Juslin and Persson's (2002) version of Medin and Schaffer's (1978) context model extended for quantitative estimates. For simplicity, the model assumes that a probe vector $\vec{x}$ is matched to all exemplar vectors $\vec{y_j}$ in memory. The criterion value c' of the probe vector is estimated as the similarity-weighted average of all n exemplar criterion values $c(\vec{y_j})$ in memory according to Equation 2.

$$c' = \frac{\sum_{j=1}^{n} S(\vec{x}, \vec{y_j}) c(\vec{y_j})}{\sum_{j=1}^{n} S(\vec{x}, \vec{y_j})} \tag{2}$$

The similarity $S(\vec{x}, \vec{y_j})$ between the probe vector $\vec{x}$ and an exemplar vector $\vec{y_j}$ is determined according to Equation 3, where D is the number of features/cues of each object:

$$S(\vec{x}, \vec{y_j}) = \prod_{i=1}^{D} d_i \quad with \quad d_i = \begin{cases} 1 \; if \; x_i = y_i \\ s_i \; if \; x_i \neq y_i \end{cases} \tag{3}$$

The weight parameter s_i determines how strong a mismatch of objects on cue i influences the similarity S that can vary between 0 and 1. For simplicity and identifiability of the model, we assume the s_i to be constant across cues.

https://doi.org/10.1017/S1930297500006513 Published online by Cambridge University Press