Judgment and Decision Making, Vol. 5, No. 5, August 2010, pp. 326–338

Cue integration vs. exemplar-based reasoning in multi-attribute decisions from memory: A matter of cue representation

Arndt Bröder*[1,2], Ben R. Newell[3], and Christine Platzer[1]

[1] University of Bonn

[2] Max Planck Institute for Research on Collective Goods, Bonn, Germany

[3] University of New South Wales, Sydney, Australia

Abstract

Inferences about target variables can be achieved by deliberate integration of probabilistic cues or by retrieving similar cue-patterns (exemplars) from memory. In tasks with cue information presented in on-screen displays, rule-based strategies tend to dominate unless the abstraction of cue-target relations is unfeasible. This dominance has also been demonstrated — surprisingly — in experiments that demanded the retrieval of cue values from memory (M. Persson & J. Rieskamp, 2009). In three modified replications involving a fictitious disease, binary cue values were represented either by alternative symptoms (e.g., *fever* vs. *hypothermia*) or by symptom presence vs. absence (e.g., *fever* vs. *no fever*). The former representation might hinder cue abstraction. The cues were predictive of the severity of the disease, and participants had to infer in each trial who of two patients was sicker. Both experiments replicated the rule-dominance with present-absent cues but yielded higher percentages of exemplar-based strategies with alternative cues. The experiments demonstrate that a change in cue representation may induce a dramatic shift from rule-based to exemplar-based reasoning in formally identical tasks.

Keywords: decision making, exemplar memory.

1 Introduction

In making choices between objects people express either preferences (Which bike do I like more?) or inferences (Which share will fare better?). Often, multiple pieces of information about attributes or probabilistic cues have to be combined. Traditionally, decision researchers formulate decision strategies as processing steps that somehow integrate the cues, either in a weighted additive fashion (Brehmer, 1994), or according to noncompensatory rules like lexicographic orderings (Gigerenzer & Goldstein, 1996). Numerous strategies have been proposed, and participants appear to choose between them adaptively (Payne, Bettman & Johnson, 1993).

Juslin, Olsson, and Olsson (2003) emphasized the structural similarity of choice and categorization tasks which both involve the integration of features (= cues). However, in categorization, exemplar-based models assuming the storage and retrieval of feature patterns as a basis for inference, rather than piecemeal cue integration (cue abstraction models, CAM), have proven successful (Brooks, 1978; Medin & Schaffer, 1978; Nosofsky, 1984). Juslin and his colleagues explored the applicability of exemplar models to multiple cue judgment tasks and found successes as well as failures (see section 1.1.3.; see also, e.g., Rieskamp & Otto, 2006; von Helversen & Rieskamp, 2009).

Persson and Rieskamp (2009; hereafter P&R) extended Juslin's approach to memory-based decisions in which cue values had to be retrieved from memory rather than being presented by the experimenter. I.e, during the judgmental phase of the experiments, participants received only the stimulus names, and all of their respective attributes — which had been learned beforehand — had to be retrieved from long term memory. To their surprise, P&R did *not* find more exemplar-based decision making; rather, most participants adopted CAM. We will test the conjecture that retrieval from memory *per se* does not induce exemplar-based decisions, whereas the difficulty of cue abstraction in combination with memory retrieval does.

*The work was supported by the Deutsche Forschungsgemeinschaft (DFG), grant BR 2130/5–1 provided to the first author. Arndt Bröder and Christine Platzer are now at the University of Mannheim. We thank Nina Bruziks, Johanna Gechter, Alexander Gehrmann, Charlotte Hartmann, Viktoriya Maydych, Matthias Mühlenberg, Dominik Pöppl, Stella Rothuysen, Helen Sauer, and Susanne Schütte for their help with the data collection in Experiment 1 or 3. Address: Arndt Bröder, University of Mannheim, Chair of General Psychology, Schloss EO, D-68131 Mannheim, Germany. Email: broeder@uni-mannheim.de.

1.1 Exemplars versus cue abstraction

To solve memory-based multi-attribute decision tasks, people can either apply CAM or exemplar models, both implying completely different kinds of knowledge representation.

1.1.1 The cue abstraction model (CAM)

Gigerenzer and Todd (1999) answered the question of how people select decision strategies with reference to the metaphor of an adaptive toolbox. Like craftsmen choose the right tools to solve specific technical problems, decision makers are assumed to choose between different decision strategies adaptively. In this context adaptivity refers to the fit between a strategy and the given environmental conditions. Therefore a strategy can never be good or bad per se but only with regard to the structure of the task environment, thus referring to the notion of bounded rationality expressed by Herbert Simon (1956). Several simple heuristics were proposed to solve decision problems and simultaneously accommodate bounded cognitive processing capacities, like the ignorance-based recognition heuristic or heuristics being referred to as one reason decision making like "Take The Best" (TTB), "Take The Last" or "Minimalist", to name just a few (see, e.g., Todd, 2001 for a classification of decision heuristics). Numerous studies show that people do select simple heuristics when the task structure is constituted in such a way, that they can outperform more complex strategies (e.g., Bröder, 2003; Bröder & Schiffer, 2006; Payne, Bettman & Johnson, 1988; Rieskamp & Otto, 2006) or when application costs are high (e.g., Bröder, 2000; Newell & Shanks, 2003; Newell, Weston & Shanks, 2003; Payne et al., 1988; Rieskamp & Hoffrage, 1999). However some empirical evidence shows that people prefer compensatory strategies that integrate a greater amount of information, when the application of such strategies is possible (Bröder & Schiffer, 2006; Rieskamp & Otto, 2006).

The term *cue abstraction* refers to the assumed knowledge representation necessary to accomplish the rule-based integration of cues. There must be some knowledge about the bivariate covariation between cue and criterion (direction and/or size of the covariation). For example, TTB searches cues in the order of their predictive validity and hence, a validity hierarchy of cues must have been established by abstracting cue-criterion relations in some learning process.

In line with P&R we used an inference task where the decision maker has to choose the alternative out of two with the higher criterion value on the basis of four cues. Several strategies can be used to solve such an inference task. Within the scope of CAM, we focus on three strategies that rely on abstract knowledge of cue-criterion relationships to make an inference. The first strategy we address is TTB, a strategy included in the adaptive toolbox (Gigerenzer & Todd, 1999). TTB is a fast and frugal heuristic because the judgment is solely based on the most valid discriminating cue. The validity of a cue is defined as the conditional probability of choosing the alternative with the higher criterion value if the judgment is solely based on this cue and the alternative with the positive cue-value is chosen (e.g., Gigerenzer & Todd, 1999). Thus TTB searches the cues in order of validity and chooses the alternative with the positive value of the first discriminating cue.

A prominent compensatory strategy is called "Weighted Additive Rule" (WADD). WADD determines the alternative with the higher criterion value by summing up weighted cue values for each alternative and by choosing the one with the largest sum. A special case of WADD is a strategy that uses identical weights. This strategy is referred to as the "Equal Weight Rule" (EQW). EQW boils down to a simple counting strategy, where the alternative with the larger number of positive cue-values is chosen. In cases where both alternatives exhibit an equal number of positive cue-values, EQW has to guess.

Of course one can think of many other strategies to solve this kind of inference task. The set of investigated strategies can always be just a sample of all possible strategies and make no claim to be exhaustive. However P&R argue that the selected strategies cover a sufficient range of strategies, where TTB represents strategies that ignore information systematically and dispense with trade-offs and WADD represents strategies that integrate a lot of information and rely on trade-offs, with EQW as a special case that is easy to apply. Based on the fact that the predictions of other strategies that rely on cue abstraction as well are highly correlated with the predictions of one of these strategies (P&R), this set of representative strategies is assumed to be sufficient to compare CAM to exemplar models.

1.1.2 The exemplar model

Contrary to CAM, exemplar models do not assume that abstract representations of cue-criterion relationships are formed during learning. Rather, each encounter with an object is simply stored in memory. For example, Brooks (1978) convincingly showed in a series of experiments that participants used knowledge about individual exemplars to accomplish a later classification task with new transfer stimuli although they had never learned categorization explicitly. In Brooks' (1978) terminology, participants judged new stimuli "by analogy" with stored exemplars. Medin and Schaffer (1978) formalized this in their notion of similarity as defined below. Hence, ac-

https://doi.org/10.1017/S1930297500002138 Published online by Cambridge University Press

cording to exemplar models, a database with cue patterns and criterion values is generated. When a new object has to be judged, the probe is compared to the stored objects, and the estimate is a weighted average of stored criterion values in which the weights are determined by the similarity between exemplars and probe (Juslin & Persson, 2002) given in Equation (1).[1]

$$S(\bar{x},\bar{y}) = \prod_{j=1}^{D} d_j$$
$$with\ d_j = \begin{cases} 1 & if\ x_j = y_i\ (feature\ match) \\ s_j & if\ x_j \neq y_i\ (feature\ mismatch) \end{cases} \quad (1)$$

D is the number of features in the probe vector $\bar{x}$ and each exemplar vector $\bar{y}$. The s_j denote attention weights given to each feature j, and they can vary between 0 and 1. Smaller numbers mean higher attention weights, since a mismatch affects the overall similarity value to a much greater extent. Like P&R, we assume that the s_j may vary between subjects, but they are constant across the four cues for each participant. According to ProbEx (the exemplar model proposed by Juslin & Persson, 2002, which is also used here) the estimation of the criterion $c(\bar{x})$ of the probe vector $\bar{x}$ is computed by multiplying the criterion $c(\bar{y_i})$ of each retrieved exemplar i with the similarity $S(\bar{x},\bar{y_i})$ between the probe and this exemplar. The estimation of the criterion $c'(\bar{x},n)$ at iteration n is given in Equation 2:

$$c'(\bar{x},n) = \frac{\sum_{i=1}^{n} S(\bar{x},\bar{y_i})c(\bar{y_i})}{\sum_{i=1}^{n} S(\bar{x},\bar{y_i})} \quad (2)$$

The estimation procedure terminates at iteration n, where the gain in accuracy of estimate by retrieving further exemplars is beneath a threshold. This is an aspect in which our exemplar model, just like the one used by P&R, differs from the more general model ProbEx: For the sake of simplicity, the ProbEx version used here assumes that all exemplars in memory are retrieved.

A cognitive representation in terms of exemplars has the advantage that no pre-processing has to occur; i.e., no cue-criterion relations have to be extracted from learning. Rather, calculations are postponed to the time of judgment. For example, if it is unclear during learning which feature will later be the criterion, an enormous computational effort would be needed to extract all possible cue-criterion relations to make them available for later rule-based processing.

[1]This description conforms to the adaption of Medin and Schaffer's (1978) classification model to quantitative estimation as suggested by Juslin and Persson's (2002) ProbEx-model.

1.1.3 Exemplar models in judgment and decision making

Karlsson, Juslin, and Olsson (2008) summarized an extensive research program which investigates when and why participants switch from rule-based cue integration to exemplar-based reasoning in judgment tasks. The latter appears to be promoted by the use of binary as opposed to continuous criterion feedback, deterministic rather than probabilistic cue-criterion relations, multiplicative rather than additive cue-criterion relations, and random as opposed to controlled learning sequence. Altogether, the results "suggest that people have an inclination to abstract explicit representations whenever possible (a 'rule bias' ...), with exemplar memory as a backup" (Juslin et al., 2003, p.153). This concurs with Brooks' (1978, p. 194) conclusion that "if there is a very simple and salient feature that predicts category membership, then adult subjects will be strongly tempted to encapsualte it in an analytic rule." This "rule bias" can persist when it is non-optimal (e.g., nonlinear environments), after extensive training with only a few exemplars, and even following instructions to use exemplar memory (Karlsson et al., 2008; Nosofsky & Bergert, 2007). Karlsson et al. (2008) conclude that the shift in strategies appears to be an active choice rather than a stimulus-driven bottom-up process. This interpretation converges with results on selecting rule-based strategies (see Bröder & Newell, 2008).

This rule bias, however, is probably present only in situations in which explicit judgments or categorizations are requested from the participants. Whenever the knowledge acquisition is incidental or implicit, this rule bias may not exist. For example, Brooks (1978) showed that, for complex rules, classification in a later transfer task was even better when the learning task did not focus on classification at all. Here, knowledge about exemplars presumably drove the performance.

1.2 Decisions from givens vs. decisions from memory

Typical decision experiments provide participants with all attribute information and then infer people's decision rules. Gigerenzer and Todd (1999) conjectured that this practice obscures the fact that information in everyday decisions often has to be retrieved from memory with associated cognitive costs probably motivating people to use simple and frugal strategies like TTB. Since TTB makes a decision based on the most valid discriminating cue, it apparently comes with fewer processing costs than compensatory strategies like WADD or EQW which integrate all cue information.

https://doi.org/10.1017/S1930297500002138 Published online by Cambridge University Press

Gigerenzer and Todd's conjecture was tested and confirmed in several experiments showing that TTB was used more often than compensatory strategies if cue information had to be retrieved from memory (Bröder & Schiffer, 2003; 2006). Decision time analyses of these experiments suggested a sequential processing of cues (Bröder & Gaissmaier, 2007). The shift to noncompensatory processing is even more pronounced under cognitive load (Bröder & Schiffer, 2006), supporting the processing cost account. One moderating variable, however, was the representational format of the cue information: verbal cues triggered TTB whereas pictorial cue patterns fostered EQW and WADD (Bröder & Schiffer, 2003; 2006). In these experiments, exemplar-based processing was not possible because, in order to use exemplars for judgment, a sample of cue patterns from a learning phase has to be stored in memory along with their criterion values. However, criterion values were never provided in the learning phase. P&R enabled exemplar-based processing by adding a learning phase that included criterion values, but still they reported almost a complete absence of exemplar-based processing. This result is surprising since memory-based judgments were thought to be especially prone to exemplar-based inference because retrieving similar exemplars might reduce processing costs relative to cue integration in working memory. We seek an explanation for this counterintuitive result in three experimental replications.

1.3 The Persson & Rieskamp study

To test their idea, P&R used a clever methodology which bears close similarities to the experimental logic introduced by Brooks (1978) and Medin and Schaffer (1978), who also used learning stimuli and later transfer stimuli in classification tasks. In addition, however, cue information about stimuli had to be retrieved from memory in the P&R study. In a first *pattern learning phase* participants learned about symptom patterns of 13 fictitious patients suffering from a mysterious tropical fever. The patients were identified by their names, and they could have 13 different constellations of symptoms out of a set of four symptoms (see Table 1). Seven learning blocks with repeated testing ensured good knowledge of the database (cue values).

In the second *feedback training phase* participants repeatedly did full paired comparisons between six of the patients, deciding which patient was in a more severe state of the disease and receiving feedback about the correct answer (criterion values). The first study used binary feedback, the second used feedback about a continuous criterion (percentage of lethal virus load in blood). We adopted the second variant. Note that participants could either store the criterion knowledge along with the already stored pattern, or they could retrieve the symptoms and extract cue-criterion relations. In the third *decision phase* participants made inferences about illness severity for the remaining 7 patients, using their cue knowledge from Phase 1 and transferring their criterion knowledge from Phase 2. Since no feedback was provided, participants had to rely on their previously acquired knowledge, which consists either of abstracted validities of each cue or exemplars stored in long-term memory.

The item set in the training phase was constructed in such a way to fulfill two conditions: First, the choice of the item set in the training phase should ensure that the exemplar model makes predictions in the decision phase that differ from the predictions of CAM in order to be able to classify participants reliably according to the strategy they most likely used. Secondly, both the application of TTB and WADD should allow for a high proportion of correct predictions in the training phase. However, neither TTB nor WADD allowed for perfect performance since there would be one exception to each rule. For example, TTB would fail in the comparison between patterns T5 and T6 whereas it would never fail in the other paired comparisons (see Table 1). Two different rank orders of the cues allowed for the same high accuracy, which implicated that two different versions of each strategy were applicable successfully.[2] Hence P&R tested two versions of TTB, namely TTB_A and TTB_B with cue orders A, B, C, D and B, A, C, D, respectively. The same holds true for WADD where the strictly compensatory cue weights 6, 4, 3, 2 were used, either for the cue order A, B, C, D ($WADD_A$) or B, A, C, D ($WADD_B$) to predict the criterion values.

P&R assessed the fit of various strategies and were surprised to find only one in 50 participants whose data fitted the ProbEx model best whereas all others were better described by rule-based models (TTB, EQW, WADD). Hence, there was no support at all for the conjecture that memory retrieval induces exemplar-based reasoning, thus strongly supporting the notion of rule based decision making.

Before one accepts this strong conclusion, it is worth examining the task more closely. In P&R's task, any symptom was either present or absent, with the presence of a symptom signaling a more severe disease. Therefore, it was probably easy to extract the cue-criterion relations.

[2]Since the criterion was computed by summing up the cue values multiplied by the weights 8, 4, 2, 1 for the cues (symptoms) A, B, C and D, respectively, a TTB strategy using validities according to this rank order should allow 100% correct inferences. Actually cue profile T5 was an exception with regard to the computation of the criterion. For cue profile T5 the criterion value 16 was assigned. Thus TTB would make a wrong choice in trials where the cue profile T5 is to be compared to the cue profile T6. To avoid this mistake TTB could use the alternative rank order B, A, C, D for the cues. However with this rank order TTB would incorrectly prefer T4 to T6.

https://doi.org/10.1017/S1930297500002138 Published online by Cambridge University Press

Table 1: Cue patterns and hypothetical criterion values used in the experiment, adopted from Persson and Rieskamp (2009). T = pattern used in the feedback learning phase, D = pattern used in the decision phase, all patterns were learned in the pattern learning phase. "1" denotes the presence of a symptom in the presence-absence format and the critical symptom in the alternative format. "0" marks the absence of a symptom or the presence of the non-critical symptom, respectively. Criterion values were computed by summing up the cue values multiplied by the weights 8, 4, 2, 1 for symptom A, B, C, D with the exception of cue profile T5, to which the criterion value 16 was assigned.

Cue profile	Symptom A	Symptom B	Symptom C	Symptom D	Criterion value
T1	0	0	0	0	0
T2	0	0	0	1	1
T3	0	0	1	0	2
T4	0	1	0	0	4
T5	0	1	1	1	16
T6	1	0	0	0	8
D1	0	0	1	1	3
D2	0	1	0	1	5
D3	0	1	1	0	6
D4	1	0	0	1	9
D5	1	0	1	0	10
D6	1	0	1	1	11
D7	1	1	0	0	12

Even a simple symptom tallying strategy (EQW) guaranteed high success rates in the decision phase; TTB and WADD fared even better (83%). If one accepts this interpretation, P&R's results fit with Karlsson et al.'s (2008) conclusion that participants will always prefer rules if the bivariate cue-criterion relations can easily be learned.

Things should be different, however, if the *direction* of the cue was not self-evident during the two learning phases. Whereas it is obvious that "fever" is associated with more sickness than "no fever", the case is less clear if you suffer either from "fever" or from "hypothermia". In the latter case with *alternative* symptoms, cue-criterion relations might be much harder to extract, and the reliance on exemplar memory might be boosted. Hence, we hypothesize that it is not only the formal structure of the learning environment that triggers different strategies, but also the semantic content of the cues which can affect the ease of cue-criterion relation extraction. Effects of semantic embedding have been reported in multiple-cue probability learning (Adelman, 1981; Muchinsky & Dudycha, 1975) and researchers have noted that learning cue direction or "polarity" is a key component of mastering multiple-cue inference tasks (Klayman, 1988; Newell, Weston, Tunney & Shanks, 2009). Related research in category learning reveals similar effects of prior knowledge on facilitating category acquisition (e.g., Wattenmaker, Murhpy, Dewey, & Medin, 1986). The basic effect is that participants learn categories in which the empirical structure of training exemplars is consistent with prior knowledge more rapidly than when structure and knowledge are inconsistent. Rehder and Murphy's (2003) knowledge resonance model accounts for this facilitation by incorporating prior concept units in to its recurrent network. These units reflect the concepts already held by participants before exposure to the experimental environment (see also Wisniewski & Medin, 1994). We tested whether prior knowledge or semantic content of the exemplars would affect participants' judgments of disease severity by replicating P&R's studies while contrasting it with a condition that was formally identical but used alternative as opposed to presence-absence cues.

2 Experiments 1 & 2

Both experiments were almost identical so we describe them together. The procedure mirrors that described by P&R with the exception of minor details in presentation (e.g., portrait photos and pictograms). The goal was to (1) replicate P&R's results and (2) to test whether the changed cue representation would promote exemplar-based decision making. The main difference between both experiments was a different rewarding scheme: Correct responses earned the participants "points". In Exper-

https://doi.org/10.1017/S1930297500002138 Published online by Cambridge University Press

iment 1, the five best participants earned a cinema ticket, whereas in Experiment 2, points were directly converted into money (0.01 € per point). The latter payoff scheme is probably more motivating and reduces the potential impact of different risk taking strategies. Second, a more homogenous sample was used in Experiment 2.

2.1 Method

2.1.1 Participants

60 participants from various fields of study and different occupations volunteered in Experiment 1 (28 female, mean age 24.7, SD = 4.16). They were acquaintances of the experimenters and received no compensation except for the chance to win one of five cinema tickets. The more homogenous sample in Experiment 2 consisted of 40 psychology students (36 female, mean age 24.5, SD = 5.87) who received course credit and strict performance-contingent payment in addition.

2.1.2 Materials and design

We used the fictitious tropical disease task of P&R. The independent variable cue representation was varied between subjects. One group learned patients and symptoms in the presence-absence format; i.e., each patient could have any combination of up to four symptoms (e.g. fever, headache, blood pressure drop, rash). For the other group, symptoms were presented in an *alternative symptoms* format. Patients always had four symptoms, for instance fever *or* weight loss, headache *or* earache, blood pressure drop *or* tachycardia, rash *or* cough. One of the symptom alternatives was critical being associated with a more severe state of the disease. (The symptom sets were counterbalanced within both experimental groups). Table 1 denotes the 13 cue patterns used in the experiments. Note that both conditions were formally identical with respect to cue-criterion relations and differed only in the labeling of the binary cues. The six patterns marked with *T* were training patterns used in the feedback learning phase. The seven patterns identified with *D* were used in the decision phase. Each fictitious patient was identified by a portrait photo and a common German male first name. Pictures and names were randomly assigned to the cue patterns for each participant.

2.1.3 Procedure

Pattern learning phase. Like P&R we used an anticipation learning paradigm (Bröder & Schiffer, 2003; 2006) with seven learning blocks, each followed by a rewarded test. In a trial of a practice block, the portrait and name of a patient were presented along with four pairs of buttons that denoted the presence or absence of a symptom (e.g., fever vs. no fever) in one condition or the alternative symptoms in the other condition (e.g., fever vs. weight loss). Participants chose a response by clicking with the mouse and received feedback via a verbal label and a pictogram symbolizing the symptom (see Figure 1a and 1b). One patient was repeated until the symptoms were reproduced without error, and then the next patient was presented. After reproducing all symptoms of all 13 patients correctly, a test followed: All 13 patients were presented, and the symptoms had to be reproduced. Participants received feedback, and earned/lost 4 points for each correct/false response (+/– 0.01 € in Experiment 2). This cycle of practice and test blocks was repeated seven times. The order of patients was randomly determined anew in all practice and test blocks.

Feedback training phase. This phase consisted of five blocks with repeated full comparisons of the six training patterns ("T" in Table 1). Full comparison of six objects results in 15 choices. Hence, this phase consisted of 75 trials. In each trial, participants were presented with two portraits and corresponding names and had to decide which of these patients was in a more severe state of the illness. The establishment of a meaningful strategy for later transfer necessitated the retrieval of the symptom patterns from memory. Participants received feedback about the correct choice and the numerical value of the criterion variable which was denoted as the viral load in the blood expressed as percentage of the lethal dose (given in Column 6 of Table 1).

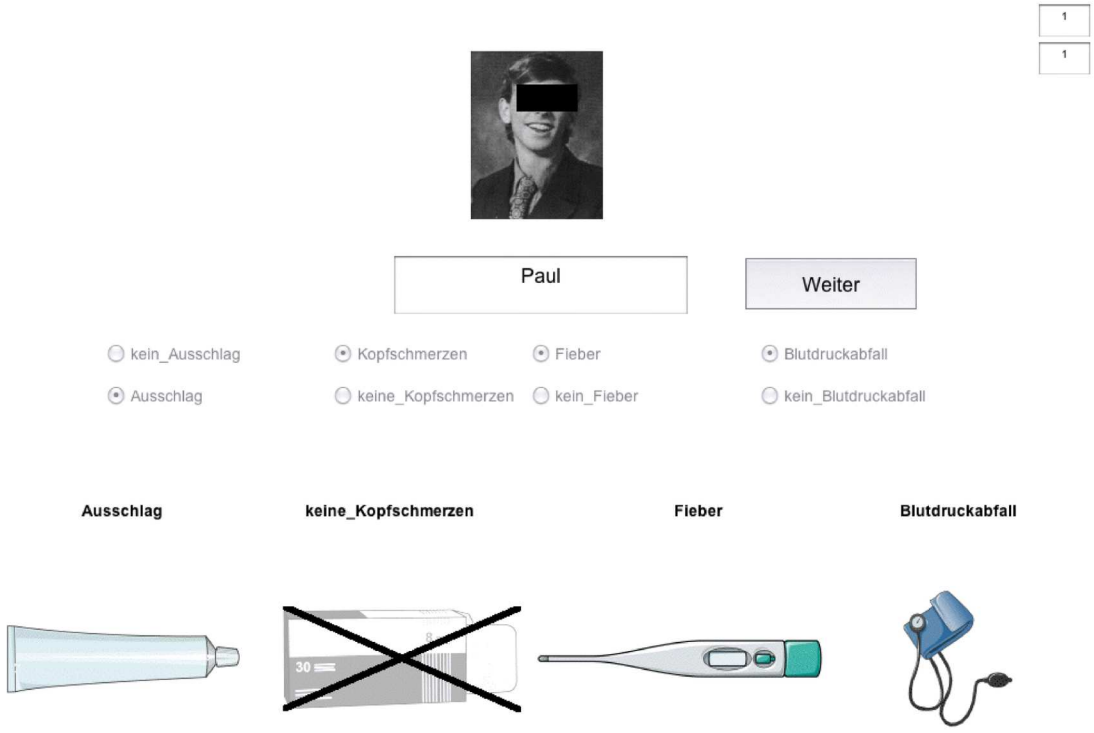
Decision phase. The third phase consisted of five blocks that contained a full set of comparisons of the remaining seven patterns *not* presented in the feedback training phase (105 trials). Participants were encouraged to use their knowledge from the former two phases and earned 15 points (0.03 € in Experiment 2) for each correct decision. However, feedback was delayed until after the decision phase was finished. This phase was crucial to compare model fits of TTB, WADD, EQW, ProbEx and Guessing and assess the strategy used.

Final memory test. Finally, participants' memory for the 13 patterns learned in the first phase was tested in the same way as in the test phases of the pattern learning phase. Each decision was sanctioned with 4/–4 points (0.03/–0.03 € in Experiment 2).

Hence, the formal structure of the task as well as almost all aspects of the procedure (except the memory test added at the end) was identical to P&R's study.

https://doi.org/10.1017/S1930297500002138 Published online by Cambridge University Press

Figure 1: a. Example of a completed pattern learning trial in the condition with presence-absence cues. (Original faces not disguised.) Ausschlag=rash, Kopfschmerzen = headache, Fieber = fever, Blutdruckabfall = blood pressure drop).



2.2 Results and discussion

2.2.1 Success of pattern learning

Figure 2 shows the learning success of the symptom patterns across the seven blocks of the first phase.

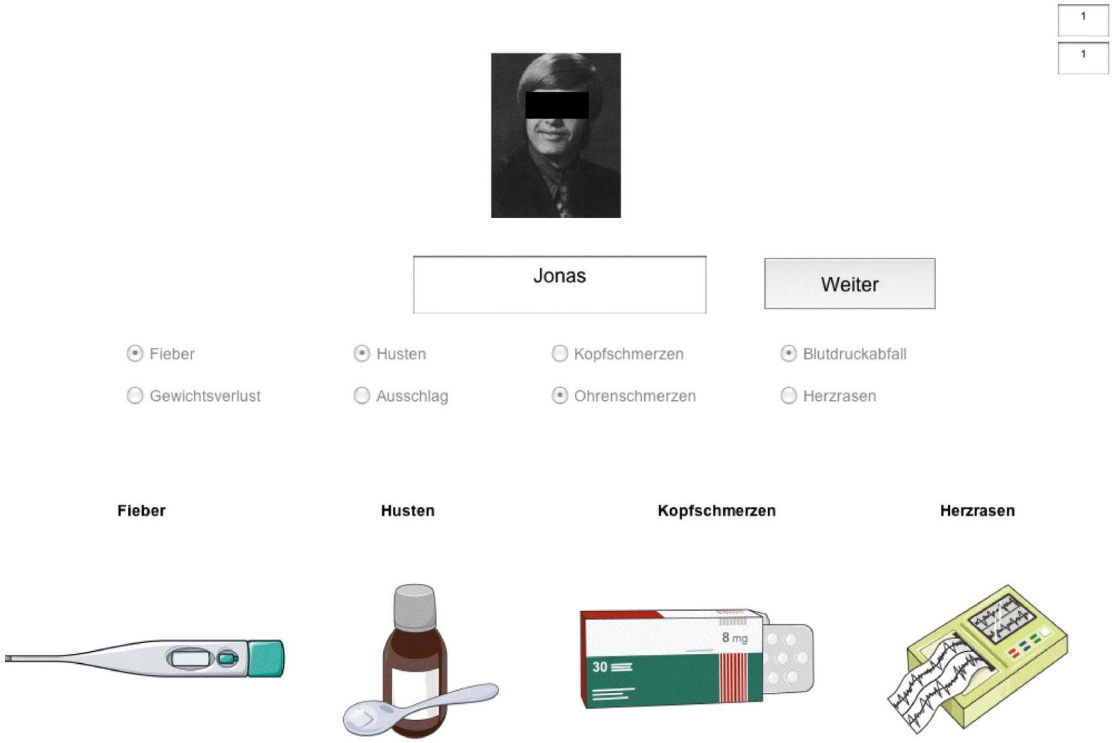
In Experiments 1 and 2, there was significant improvement across blocks (Huyn-Feldt $F(3.68, 213.35) = 76.87$, $p < .001$ and $F(2.78, 105.58) = 96.29$, $p < .001$, respectively), a main effect of the experimental condition showing better performance in the presence-absence condition ($F(1,58) = 15.23$, $p < .001$ and $F(1,38) = 8.66$, $p = .006$) as well as a tendency for an interaction, indicating slower learning in the alternative-cues conditions ($F(3.68, 213.35) = 2.59$, $p = .04$ and $F(2.78, 105.58) = 2.36$, $p = .08$). Hence, alternative cues were harder to learn, which is not surprising because they convey more verbal (but not statistical) information. In addition, Figure 2 shows, as intended, better performance in the second experiment, possibly due to the more motivating reward scheme. The final memory performance was 87% correct in the presence-absence condition and 74% in the alternative-cue condition in Experiment 1. The improved values were 97% and 83%, respectively, in Experiment 2. Hence, the cue patterns were established reasonably well in memory to warrant further analysis.

2.2.2 Decision phase

An outcome-based maximum-likelihood method was used to assess individual choice strategies. This classification method is aimed at determining the highest likelihood of the data, given precise predictions derived from each of the cognitive models. Therefore it is essential that the item set in the decision phase is designed in such a way that each strategy predicts a different choice pattern across the decision trials. By comparing the observed choice patterns of participants to the predicted choice patterns of each strategy, the best-fitting strategy can be identified and the participant is classified as user of this strategy. Assuming that participants sometimes make errors when using a decision strategy, simple binomial response error models are formulated that serve as a basis for classification. The ML method computes conditional probabilities of the observed data, given each of the strategies and response errors. Additionally this method provides likelihood ratios as a measure of confidence in the correctness of the classification (for details see e.g., Bröder

https://doi.org/10.1017/S1930297500002138 Published online by Cambridge University Press

Figure 1: b. Example of a completed pattern learning trial in the condition with alternative cues. (Original faces not disguised.) Fieber = fever, Gewichtsverlust = weight loss, Husten = cough, Ausschlag = rash, Kopfschmerzen = headache, Ohrenschmerzen = earache, Blutdruckabfall = blood pressure drop, Herzrasen = tachycardia).



& Schiffer, 2003; Bröder, 2010).[3]

Predictions of the following strategies were generated: TTB_A, TTB_B, $WADD_A$, $WADD_B$, EQW, $ProbEx_A$, and $ProbEx_B$. TTB_A and TTB_B refer to TTB strategies using different cue rankings. As P&R remarked about their environment, both variants are equally successful in the feedback training phase and hence, participants might learn different optimal cue orders. The same holds true for WADD (see section 1.3). In contrast to P&R we also implemented two different versions of ProbEx, differing with regard to their s-parameter. P&R used a restricted variant of ProbEx assuming equal weights $s_j = 0.5$ for each cue j (see P&R, for a justification). For the sake of parsimony we also assumed equal weights for each cue j, but the s-parameter was not set to a fixed value of $s = 0.5$ for every participant but was a free parameter. As mentioned above the s-parameter determines the weight, with which dissimilar exemplars contribute to the estimation of the criterion value of a certain probe. The selected item set for the decision phase had the advantage that every possible value of s led to only two different prediction vectors of ProbEx. Hence we implemented two different versions of ProbEx with $ProbEx_A$ presuming $s > 0.436$ and $ProbEx_B$ presuming $s < 0.436$.

Each participant was classified as using one of these strategies according to the best model fit score when the estimated response error $\hat{\varepsilon}$ for the best fitting model was less than 0.40. Making an error with probability $\hat{\varepsilon}$ means that in 40 percent of all trials the alternative not predicted by the strategy was erroneously chosen. If the best fitting model yielded $\hat{\varepsilon} > 0.40$, a pattern was classified as a random guessing strategy (see, e.g., Bröder & Schiffer, 2003). The Maximum Likelihood classification method computes the likelihood of the data under each strategy. Hence, one can compute pairwise Bayes factors (likelihood ratios) as measures of classification reliability. We computed the ratios of likelihoods for the best-fitting model and the second best for each participant. This Bayes factor denotes how much more likely the data are under the assumption of the best strategy than under the assumption of the second best. According to conventions that can be found in Wassermann (2000), the clas-

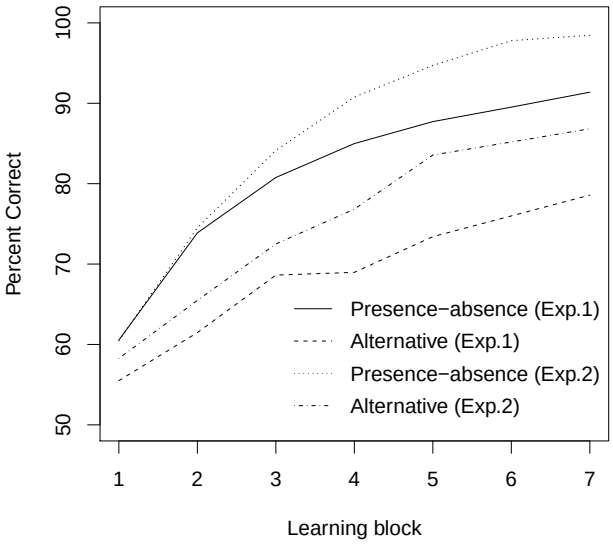
[3]P&R used a quadratic scoring rule (QS) to assess model fits and reported an average QS fit for each strategy and reported it on a group level. Group averages of a fit measure confound the actual fit of the strategy with its actual frequency in the group. The ML method provides the error rate as an easy-to-interpret fit measure per participant and strategy. The pattern of results, however, does not change, if P&R's classification method is used.

https://doi.org/10.1017/S1930297500002138 Published online by Cambridge University Press

Table 2: Likelihood Ratios according to strategy classifications across all three experiments (likelihood of strategy with most likely data divided by second largest likelihood), TTB = Take The Best, WADD = Weighted Additive Rule, EQW = Equal Weight Rule, ProbEx = exemplar model, conventions for weak / moderate / strong evidence in favour of a model after Wassermann (2000).

			Likelihood Ratio (Bayes factor)			
	N	Min	% of participants < 3 (weak evidence)	% of participants with 3 to 10 (moderate evidence)	% of participants > 10 (strong evidence)	Md
TTB	13	9.20	0	7.7	92.3	6149.74
WADD	49	1.78	14.3	10.2	75.5	902.89
EQW	41	1.38	19.5	2.4	78.0	3667.81
ProbEx	32	1.38	9.4	18.8	71.9	189.36

Figure 2: Correct reproductions of symptoms across the seven blocks in the pattern learning phase of Experiments 1 & 2.



sification of the vast majority of participants (more than 80% for each strategy) could be classified with at least moderate confidence (Bayes factor larger than 3; see Table 2.

Table 3 reports the average percentage of predicted inferences in all experimental conditions of all experiments by strategy,[4] as well as the strategy classifications in all experimental conditions. For the sake of clarity, the different versions of the strategies (A vs. B) are presented in aggregate.

[4]Note that this corresponds to $(1 - \varepsilon)$, where ε is the estimated error probability of a strategy.

Whereas no participant used ProbEx in the presence-absence cue conditions of Experiments 1 and 2, respectively, the number increased to 10 (34%) and 8 (40%) in the respective alternative cue conditions. Contrasting the frequency distributions between conditions yielded significant differences in both experiments ($\chi^2(4) = 17.73$ and $\chi^2(4) = 18.10$, both $p < .01$).

Hence, we replicated P&R's result that ProbEx apparently plays no role in memory-based decisions. This was true, however, only for a presence-absence cue format. With an alternative cue format, the proportion of ProbEx users increased up to 40%. Our results thus corroborate our hypothesis that binary cues with distinct alternatives trigger exemplar-based inference. We hypothesize that this format increases the difficulty of cue abstraction during training since not only cue validity orders have to be determined, but also the cue directions.

Two major experimental confounds may possibly undermine this conclusion in the first 2 experiments: First, as one reviewer acknowledged, participants might have interpreted the four binary alternative symptoms as eight independent symptoms. Although we consider this possibility quite unlikely, given the instructions, the display during pattern learning, and the lack of co-occurrences of exclusive symptoms, we conducted a third experiment in which we ruled out this possibility of a misrepresentation. Second, the final memory performance differed between groups because alternative symptoms were harder to learn. Strategy differences might therefore reflect the quality of the memory representation rather than its nature. This is a serious caveat. In the first two experiments, we used 7 learning blocks for both groups in order to match P&R's procedure as closely as possible. In our third experiment, all participants learned the pattern to the same success criterion in order to eliminate differences in cue knowledge.

https://doi.org/10.1017/S1930297500002138 Published online by Cambridge University Press

Table 3: Frequencies and average percentage of predicted inferences of strategies used, classified using a ML estimation according to the best-fitting model, Chi-square values contrast ProbEx vs. CAM across experimental conditions. TTB = take the best, WADD = weighted additive strategy, EQW = equal weight strategy, ProbEx = exemplar model, Guess = guessing (percentage of predicted inferences < 60%), Unclass. = unclassified pattern (identical likelihoods for 2 strategies).

		Strategy classification						
		TTB	WADD	EQW	ProbEx	Guess	Unclass.	
Exp. 1	presence-absence	4 82.14%	13 87.66%	10 94.33%	- -	4 -	- -	$\chi^2(1) = 16.24$, $p < .001$
	alternative	1 86.67%	5 74.70%	5 75.33%	10 66.86%	7 -	1 -	
Exp. 2	presence-absence	3 94.60%	11 93.85%	5 92.67%	- -	1 -	- -	$\chi^2(1) = 14.33$, $p < .001$
	alternative	2 85.24%	3 78.41%	1 80.00%	8 66.79%	5 -	1 -	
Exp. 3	presence-absence	2 87.14%	12 87.14%	14 94.29%	2 74.76%	- -	- -	$\chi^2(1) = 13.04$, $p < .001$
	alternative	1 71.43%	5 77.34%	6 90.00%	12 69.52%	8 -	- -	

3 Experiment 3

In the third experiment, both confounds were eliminated: Cue labels in the alternative condition were mutually exclusive symptoms (fever vs. hypothermia; constipation vs. diarrhea; hepatomegaly vs. cirrhosis; hypertension vs. hypotension), and all participants were required to meet a 90% learning criterion in the learning phase in a maximum of 15 learning blocks. Participants did not earn or lose points in the pattern learning phase. In the instructions, all symptoms were explained, so their pairwise exclusive nature was obvious to the participants. Experiment 3 resembled the former experiments in all other respects.

3.1 Participants

62 people, mainly psychology students participated in Experiment 3 (55 female, mean age = 21.58, SD=3.64). The participants received course credit and performance contingent payment with the best 40% of participants earning additional 10 €.

3.2 Results and discussion

3.2.1 Success of pattern learning

Figure 3 shows the cumulative percentage of participants who reached the learning criterion within a certain learning block.

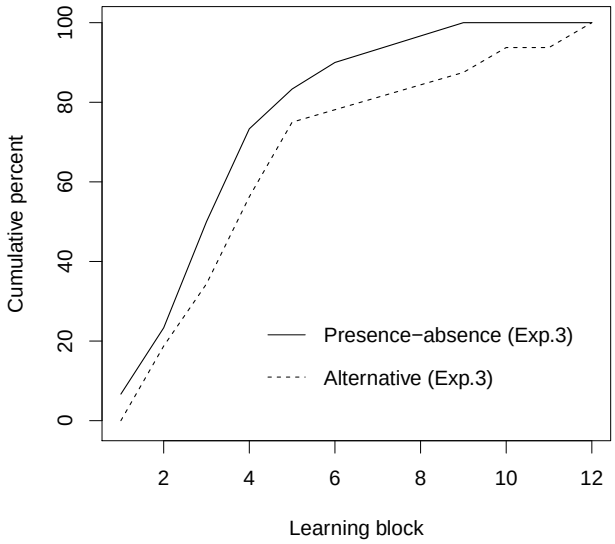
Whereas 100 percent of the participants in the presence-absence condition reached the learning criterion within 9 learning blocks, it took 12 learning blocks for the participants in the alternative condition. This difference was not significant ($t(60) = 1.81$, $p > .05$).

3.2.2 Decision phase

Strategy classifications and fit values are provided in Table 3. There were 2 (6%) versus 12 (39%) ProbEx users in the presence-absence condition and alternative condition, respectively ($\chi^2(4) = 21.52$, $p < .001$), confirming the former results. Since memory performance in the final test was equivalent between groups (92.18% vs. 91.53%, $t(60) = 0.54$, $p =. 59$) and the mutually exclusive symptom labels excluded the eight cues interpretation, the confirmation of the general result cannot be attributed to these possible confounds being present in the first two experiments.

https://doi.org/10.1017/S1930297500002138 Published online by Cambridge University Press

Figure 3: Cumulative percentage of participants who reached the learning criterion (90 percent correct reproductions of symptoms) in a certain learning block in Experiment 3.



4 General discussion

Karlsson et al. (2008) reviewed an extensive research program showing that exemplar-based reasoning belongs to the "toolbox" of strategies in multi-attribute decision making but that people generally exhibit a preference for strategies based on cue abstraction. Only if the abstraction of cue-criterion relations becomes hard do people switch to exemplar-based strategies. Sometimes, they are even reluctant to do so when explicitly instructed to use exemplar memory (Nosofsky & Bergert, 2007). In summary, Juslin and co-workers focused on formal characteristics of environments that trigger exemplar-based inferences.

P&R extended this view in hypothesizing that the need to retrieve cue and exemplar information from memory might also foster a shift to exemplar reasoning. To their surprise, there was no such tendency. We added another psychological hypothesis, stating that memory-based decisions *per se* do not necessarily induce exemplar-based reasoning when cue-criterion relations are still easy to extract. Furthermore, cue abstraction is not only influenced by the formal structure of the environment, but also by the cognitive representation of cues. Remember that the formal structure of both conditions was identical in our experiments.

Hence, *neither* memory-based decisions nor an alternative cue format *per se* are sufficient for triggering exemplar judgments. Note that the procedure used by Juslin and his co-workers always involved cues with alternative cue values in which new cue patterns are presented visually during both the feedback learning phase and the decision phase. Hence, cue criterion relations could be learned without burdening working memory. Not until other factors are supervened (e.g., binary, probabilistic or multiplicative criterion) that complicate the abstraction of cue-criterion relations, people switch to exemplar models as a backup. Newell et al. (2009) found that participants' learning in tasks similar to those used by Juslin, Olsson et al. (2003) depended crucially on the ability to infer cue direction — a finding that echoes the difference in the learning performance of the presence-absence and alternative cues groups in the current experiments (see Figure 2). In a recent paper, von Helversen and Rieskamp (2009) also explored the predictive accuracy of exemplar models. In line with the previously mentioned results, they found a preponderance of exemplar-based reasoning only in a condition with an alternative cue format with unknown cue direction in combination with a reduced number of predictive cues (only 3 out of 6 cues correlated substantially with the criterion). From these findings one could conclude that the crucial factor for triggering exemplar-based processes is the availability of knowledge concerning the task structure, namely knowledge that is indispensable for inferring the underlying rules. The fact that the task is memory-based may be of secondary importance. However we would argue that it is the interaction of different factors making exemplar based reasoning necessary because, due to a cognitive overload, rules can no longer be inferred. The nature and number of factors interacting can be manifold. Providing alternative cues with an unknown cue direction does not per se trigger exemplar based processes but only in combination with a binary or deterministic criterion (Juslin, Olsson et al., 2003), a multiplicative cue-combination rule (Juslin, Karlsson & Olsson, 2008), a multiplicative rule plus a reduced number of predictive cues (von Helversen & Rieskamp, 2009) or if cue information has to be retrieved from memory (present study).

Our results clearly corroborate the hypothesis of a strategy shift from rule-based to exemplar-based reasoning with an alternative cue format. However one aspect that is conspicuous and needs further explanation is that the percentage of predicted inferences for ProbEx is less in both conditions compared to the percentage of predicted inferences for CAM. This can be interpreted as a higher proportion of unsystematic response errors, for ProbEx than for CAM. This finding seems to be plausible at least for two reasons, albeit the explanation is post hoc: As mentioned above, Juslin and colleagues conceptualize ProbEx as a backup whenever cue abstraction is not possible (Juslin, Olsson et al., 2003). Abstract rule-based knowledge has the advantage that "[... it] is explicit, can be verbalized, and therefore is likely to create a stronger sense of insight into and mastery of the task"

https://doi.org/10.1017/S1930297500002138 Published online by Cambridge University Press

(Juslin, Jones, Olsson & Winman, 2003, p. 938). Furthermore abstract rules can be generalized more easily and provide estimations of the criterion values that are more robust than the ones provided by the exemplar model. In the present experiments cue pattern T5 is an example of an exemplar with an extreme criterion value (see Table 1). Such extreme exemplars have generally less influence on the abstraction of a rule. However the estimation of the criterion value is much more distorted if such exemplars serve as a basis for exemplar based reasoning. Maybe people go against this influence intuitively by considering extreme exemplars less important, thus producing a response pattern that is more dissimilar to the one predicted by ProbEx. Another explanation focuses on the time, when computations are required: CAM has to abstract validities during the feedback training phase and during the decision phase no further computations are required. In contrast ProbEx postpones all computations to the time of the judgment. During the decision phase ProbEx has to compute and integrate similarities between the probe and stored exemplars. For this reason, the probability of making unsystematic response errors in the decision phase is higher for ProbEx than for CAM (see also Juslin & Persson, 2002). Finally, the process of computing a probe's similarity to stored exemplars may simply be noisier than rule-based cue integration and thus lead to noisier responses. This may be the very reason why people show the "rule bias" and use exemplar-based reasoning only as a backup in probabilistic environments.

What the current study adds is new insight into how the learning of cue direction is affected by the need to *retrieve* cue information in each learning trial and keep it temporarily available in working memory. With presence-absence cues, it is still manageable to extract the cue-criterion relation since the presence of a symptom always points to a more severe state of the sickness. The dominance of the WADD strategy in the presence-absence condition across all three experiments attests to the ease with which participants presumably added up such cues when drawing inferences about disease severity. With alternative cues, however, additional memory processes are necessary: You also have to retrieve the direction of the relationship (i.e., which symptom is associated with more severe illness? which one is not?). Our results suggest that only the combination of both factors (memory retrieval and alternative cues) burdens working memory enough to have many people switch to similarity-based processing.

References

Adelman, L. (1981). The influence of formal, substantive, and contextual task properties on the relative effectiveness of different forms of feedback in multiple-cue probability learning tasks. *Organizational Behavior & Human Performance, 27*, 423–442.

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

Bröder, A. (2000). Assessing the empirical validity of the "Take The Best" heuristic as a model of human probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 1332–1346.

Bröder, A. (2003). Decision making with the "adaptive toolbox": Influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 611–625.

Bröder, A. (2010). Outcome-based strategy classification. In A. Glöckner, & C. L. M. Witteman (Eds.), *Foundations for tracing intuition: Challenges and methods* (pp. 61–82). London: Psychology Press & Routledge.

Bröder, A., & Gaissmaier, W. (2007). Sequential processing of cues in memory-based multiattribute decisions. *Psychonomic Bulletin & Review, 14*, 895–900.

Bröder, A., & Newell, B.R. (2008). Challenging some common beliefs about cognitive costs: Empirical work within the adaptive toolbox metaphor. *Judgment and Decision Making, 3,* 195–204.

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.

Bröder, A., & Schiffer, S. (2006). Stimulus Format and Working Memory in Fast and Frugal Strategy Selection. *Journal of Behavioral Decision Making, 19*, 361–380.

Brooks, L. R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (eds.), *Cognition and Categorization* (pp. 169–215), Hillsdale, NJ: Erlbaum.

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

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: the adaptive toolbox. In G. Gigerenzer, P. M. Todd & the ABC Research Group, *Simple heuristics that make us smart* (pp. 3–34). New York: Oxford University Press.

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

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division

of labor hypothesis. *Cognition, 106,* 259–298.

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

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

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.

Klayman, J. (1988). On the how and why (not) of learning from outcomes. In B. Brehmer & C. R. B. Joyce (Eds.), *Human Judgment: The SJT view* (pp. 115–160). North-Holland: Elsevier.

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

Muchinsky, P. M., & Dudycha, A. L. (1975). Human inference behavior in abstract and meaningful environments. *Organizational Behavior & Human Performance, 13*, 377–391.

Newell, B. R., & Shanks, D. R. (2003). Take the best or look at the rest? Factors influencing one-reasen decision-making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 53–65.

Newell, B. R., Weston, N. J., & Shanks, D. R. (2003). Empirical tests of a Fast-and-Frugal heuristic: Not everyone "takes-the-best". *Organizational Behavior & Human Decision Process, 91,* 82–96.

Newell, B. R., Weston, N. J., Tunney, R. J., & Shanks, D. R. (2009). The effectiveness of feedback in multiple-cue probability learning. *The Quarterly Journal of Experimental Psychology, 62*, 890–908.

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

Nosofsky, R. M., & Bergert, F. B. (2007). Limitations of exemplar models of multi-attribute probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 999–1019.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 534–552.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker.* Cambridge: Cambridge University Press.

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

Rehder, B., & Murphy, G. L. (2003). A knowledge-resonance (KRES) model of category learning. *Psychonomic Bulletin & Review, 10*, 759–784.

Rieskamp, M., & Hoffrage, U. (1999). When do people use simple heuristics, and how can we tell? In G. Gigerenzer, P. M. Todd & the ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 141–167). New York: Oxford University Press.

Rieskamp, J., & Otto, P. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General, 135,* 207–236.

Simon, H. A. (1956). Rational choice and structure of environments.*Psychological Review, 63,* 129–138.

Todd, P. M. (2001). Fast and frugal heuristics for environmentally bounded minds. In G. Gigerenzer & R. Selten (Eds.), *Bounded Rationality. The adaptive toolbox* (pp. 51–70). Cambridge, MA: The MIT Press.

Von Helversen, B., & Rieskamp, J. (2009). Models of Quantitative Estimations: Rule-Based and Exemplar-Based Processes Compared. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 867–889.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology, 44*, 92–107.

Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. L. (1986). Linear separability and concept learning: Context, relational properties, and concept naturalness. *Cognitive Psychology, 18*, 158–194.

Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science, 18*, 221–282.