Cue Abstraction and Exemplars in Multiple-Cue Judgment

Peter Juslin (peter.juslin@psy.umu.se)
Department of Psychology, Umeå University
SE-901 87, Umeå, Sweden

Henrik Olsson (henrik.olsson@psy.umu.se)
Department of Psychology, Umeå University
SE-901 87, Umeå, Sweden

Anna-Carin Olsson (anna-carin.olsson@psy.umu.se)
Department of Psychology, Umeå University
SE-901 87, Umeå, Sweden

Abstract

Although categorization and multiple-cue judgment are similar tasks, categorization models emphasize *exemplar memory*, while multiple cue judgment routinely is interpreted in terms of mental integration of cue weights that are abstracted in training. We investigate if these conclusions derive from genuine differences in the processes in the two tasks or are accidental to different research methods. The results reveal large individual differences and a shift from exemplar memory to mental cue-abstraction when the criterion is changed from classification to continuous. This suggests that people switch between qualitatively distinct processes in the two tasks.

Introduction

A categorization task typically requires a probe described by a number of binary *features* to be classified into one, of usually two, *categories*. A multiple-cue judgment involves a probe defined by binary or continuous *cues* and typically requires judgment of a continuous *criterion*. Both tasks require inference from known variables to an unknown variable. Despite the structural similarity of the tasks (Figure 1), the most successful cognitive models in the two domains are profoundly different in terms of the computations, cognitive processing, and neural substrate that they imply. Research on categorization often emphasize *exemplar memory* (e.g., Nosofsky & Johansen, 2000): retrieval of memory traces of concrete objects from different categories. In research on multiple-cue judgment, the (explicit or implicit) interpretation is generally that people retrieve abstracted knowledge of cue weights, which is then mentally integrated to perform a judgment (e.g., Einhorn, Kleinmuntz, & Kleinmuntz, 1979).

In this article, we report an investigation into the reasons for these divergent conclusions. From the outset, we can identify two possible answers. The first is that research on multiple-cue judgment has not benefited from the designs and the cognitive modeling needed to disclose the importance of exemplar memory. From this point of view, the conclusions are accidental to different research paradigms and once that we scrutinize the processes carefully we find that they are essentially the same. A second answer is that the different conclusions derive from the differences that nonetheless distinguish the two tasks; for example, the use of a binary criterion in categorization tasks and a continuous criterion in multiple cue judgment tasks. The latter answer suggests a cognitive system with multiple levels of qualitatively distinct representations that compete to control behavior depending on the requirements (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Jones, Juslin, Olsson, & Winman, 2000; Juslin, Olsson, & Olsson, 2002).

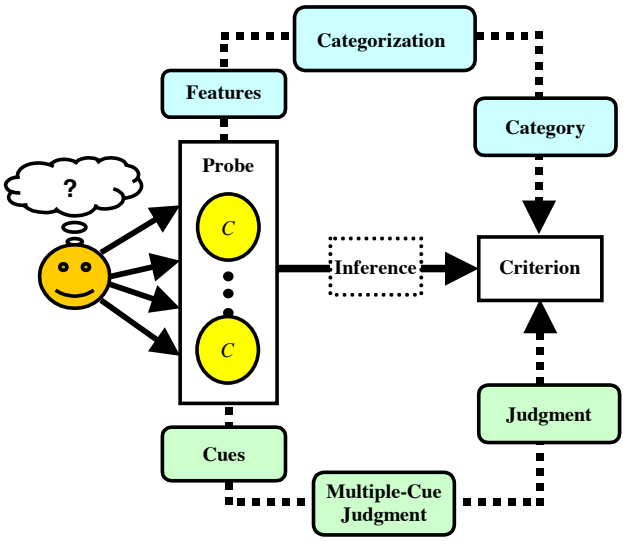

Figure 1: The structural similarity between a categorization task and a multiple-cue judgment task.

The Judgment Task

The task requires participants to use four binary cues to infer a binary or continuous criterion (Jones et al., 2000; Juslin et al., 2002). The judgments involve the toxicity of subspecies of an exotic (but fictitious) Death Bug. The different subspecies vary in concentration of poison from 50 ppm to 60 ppm (a continuous criterion), where concentrations below 55 ppm are harmless but concentrations above 55 ppm are lethal (a binary criterion, harmless vs. dangerous). The toxicity can be in-

ferred from four binary cues of the subspecies (e.g., short or long legs, spots or no spots on the fore-back).

The cues take on values 1 or 0 and the toxicity c of a subspecies is a linear, additive function of the cues:

$$c = 50 + 4 \cdot C_1 + 3 \cdot C_2 + 2 \cdot C_3 + 1 \cdot C_4 . \qquad (1)$$

C_1 is the most important cue with *coefficient* 4 (i.e., a relative weight .4), C_2 is the second to most important with coefficient 3, and so forth. The binary criterion b is formed from the continuous criterion by assigning c<55 b=0 (harmless), c>55 b=1 (dangerous), and c=55 randomly as b=1 or b=0. A subspecies with feature vector (0, 0, 0, 0) thus has poison concentration 50 ppm and is harmless; a subspecies with feature vector (1, 1, 1, 1) has 60 ppm and is dangerous. The continuous and the binary criteria for all the 16 subspecies (i.e., possible cue configurations) are summarized in Table 1.

In *training*, the participants encounter 11 subspecies and make either *binary judgments* about the toxicity of each subspecies (i.e., "harmless" or "dangerous") or *continuous judgments* about their toxicity (e.g., "The amount of poison is 57 ppm"). As indicated in the two right-most columns of Table 1, five subspecies are omitted in training. (Sets A and B, respectively, denote two different training sets where three omitted subspecies are counter-balanced.) In a test phase, the participants make the same judgments as in the training phase, but for all 16 subspecies and without feedback.

Table 1: Structure of the judgment task. The out-balanced constrained training sets are denoted A and B.

Exemplar	Cues				Criteria		Set	
#	C_1	C_2	C_3	C_4	Cont.	Bin.	A	B
1	1	1	1	1	60	1	E	E
2	1	1	1	0	59	1	T	T
3	1	1	0	1	58	1	T	T
4	1	1	0	0	57	1	O	N
5	1	0	1	1	57	1	N	O
6	1	0	1	0	56	1	N	O
7	1	0	0	1	55	p=.5	N	O
8	1	0	0	0	54	0	T	T
9	0	1	1	1	56	1	O	N
10	0	1	1	0	55	p=.5	O	N
11	0	1	0	1	54	0	T	T
12	0	1	0	0	53	0	T	T
13	0	0	1	1	53	0	T	T
14	0	0	1	0	52	0	T	T
15	0	0	0	1	51	0	T	T
16	0	0	0	0	50	0	E	E

Note: E = Extrapolation exemplar, T = training exemplar, O = Old comparison exemplar presented in training, matched on the criterion to one of the new exemplars, N = New comparison exemplar presented the first time at test, p=.5 assigns binary criterion 1 to the exemplar with probability .5.

A criticism of previous studies that support exemplar models is that often the artificial categories used essentially contain no structure at all. There is thus, in a sense, no other way to solve the task than to memorize the exemplars (Smith & Minda, 2000). Our task is neutral in this respect because it allows perfect performance in training both by exemplar memory and by induction of the task structure (i.e., by inducing Eq. 1).

Cognitive Models

The *cue-abstraction model* assumes that participants abstract explicit cue-criterion relations in training which are mentally integrated at the time of judgment. When presented with a probe the participants retrieve rules connecting cues to the criterion from memory (e.g., "Green back goes with being poisonous"). The rules specify the sign of the contingency and the importance of the cue with a cue weight. For example, after training the rule for cue C_1 may specify that C_1=1 goes with a large increase in the toxicity of a subspecies.

With a continuous criterion, cue abstraction suggests that the participants compute an estimate of the continuous criterion c. For each cue, the appropriate rule is retrieved and the estimate of c is adjusted according to the cue weight ω_i (i=1…4). The final estimate $\hat{c}_R$ of c is a linear additive function of the cue values C_i,

$$\hat{c}_R = k + \sum_{i=1}^{4} \omega_i \cdot C_i , \qquad (2)$$

where k = 50 + .5(10-$\sum\omega_i$). If ω_1 =4, ω_2=3, ω_3=2, and ω_4=1, Eq's 1 and 2 are identical and the model produces perfect judgments. The intercept k constrains the function relating judgments to criteria to be regressive around the midpoint (55) of the interval [50, 60] specified by the task instructions[1]. This formulation essentially provides a cognitive interpretation of the linear additive model known to provide a good account of multiple-cue judgment data (Brehmer, 1994). Predictions by the cue-abstraction model in a continuous task are illustrated in Figure 2A.

The binary judgment involves classification of subspecies into two categories based on their continuous criterion. One way to obtain such judgments from Eq. 2 is by assigning all subspecies with $\hat{c}_R$<.5 as harmless and all subspecies with $\hat{c}_R$>.5 as dangerous. Whenever the estimates are correct ($\hat{c}_R$=c) this implies a relation between classification proportions $p_R(b=1)$ and the

[1] The constrained formulation captures the regression effect within the interval [50, 60] that is introduced by a random error in the cue weights or the process of cue abstraction. For example, for the extreme subspecies, (0, 0, 0, 0: c=50) and (1, 1, 1, 1: c=60), random error may produce judgments that deviate from 50 and 60, respectively. However, for exemplar (0, 0, 0, 0: c=50) we expect the errors to more often produce a judgment above than below 50. For exemplar (1, 1, 1, 1: c=60) we expect the errors to more often produce a judgment that is below than above 60. Second: it holds to a good approximation in the data reported below. Third, it provides a four-parameter implementation that is more easily compared to the four-parameter exemplar model described below in terms of the number of free parameters.

criterion c that is a step function. Taking into account that the process is likely to involve error in cue abstraction and decision making, we allow for a sigmoid function in the form of a logistic function (see Figure 1A):

$$\hat{p}_R(b=1) = \frac{e^{k+\sum W_i \cdot C_i}}{1+e^{k+\sum W_i \cdot C_i}}. \qquad (3)$$

where W_i are the cue weights in a logistic regression and k = -.5 $\sum W_i$. The intercept k implies a crossover from binary judgment 0 to b 1 at toxicity 55, as implied by the instructions. When the cue-abstraction model is fitted to binary judgments below, we rely on Eq. 3.

Exemplar models suggest that the participants make judgments by retrieving similar exemplars (subspecies) from long-term memory. The *context model* of perceptual classification (Medin & Schaffer, 1978) suggests that the probability $\hat{p}_E(b=1)$ of categorization as dangerous equals the ratio between the summed similarity of the judgment probe to the dangerous exemplars and the summed similarity to all exemplars:

$$\hat{p}_E(b=1) = \frac{\sum_{j=1}^{J} S(p,x_j)\cdot b_j}{\sum_{j=1}^{J} S(p,x_j)}, \qquad (4)$$

where p is the probe to be judged, x_j is stored exemplar j (j=1…J), $S(p,x_j)$ is the similarity between the probe p and exemplar x_j, and b_j is the binary criterion stored with exemplar j (b_j=1 for dangerous, b_j=0 for harmless). J depends on the size of training set of exemplars.

The similarity between probe p and exemplar x_j is computed by the multiplicative similarity rule of the context model (Medin & Schaffer, 1978):

$$S(p,x_j) = \prod_{i=1}^{4} d_i, \qquad (5)$$

where d_i is an index that takes value 1 if the cue values on cue dimension i coincide (i.e., both are 0 or both are 1), and s_i if they deviate (i.e., one is 0, the other is 1). s_i are four parameters in the interval [0, 1] that capture the impact of deviating cues (features) on the overall perceived similarity $S(p,x_j)$. s_i close to 1 implies that a deviating feature on this cue dimension has no impact and is considered irrelevant. s_i close to 0 means that the overall similarity $S(p,x_j)$ is close to 0 if this feature is deviating, assigning crucial importance to the feature. The parameters s_i capture the similarity relations between stimuli and the attention paid to each cue dimension, where a lower s_i signifies higher attention.

The context model was developed for classification. To generate predictions also for judgments of a continuous criterion we relax the model by allowing the outcome index b_j to be a continuous value. The estimate $\hat{c}_E$ of c is then a weighted average of the criteria c_j stored for the exemplars, with similarity $S(p,x_j)$ as the weights (see e.g., DeLosh, Busemeyer, & McDaniel, 1997; Juslin & Persson, 2000; Smith & Zarate, 1992).

Predictions

The predictions are summarized in Figures 2 (binary criterion) and 3 (continuous criterion). In both tasks, the models produce similar predictions when all exemplars are presented both at training and test (the upper panels). Both models thus provide accurate representations of the environment, albeit by different means. Figures 3A and 3B illustrate that the good fit of a linear additive model need not be informative in regard to whether cues are really mentally integrated according to a linear model: predictions by an exemplar model are identical. When the extreme exemplars (c= 50 & 60) and three intermediate exemplars (c=.55, 56, & 57) are withheld in training, the models produce distinct predictions.

As illustrated in the lower panels of Figures 2 and 3, the cue abstraction model allows accurate extrapolation beyond the distribution of criteria in the training set [51, 59]. Whenever the correct signs of the cue weights are identified, the most extreme judgments are made for exemplars 1 (c=60) and 16 (c=50). The exemplar model that computes a weighted average of the criteria observed in training can never produce a judgment outside the observed range (Delosh et al., 1997). The most extreme judgments are made for criteria c=51 and 59.

With the cue abstraction model there should be no systematic difference between judgments for the "New" and "Old" exemplars with c=55, 56, and 57: the process is essentially the same in both cases. However, with the exemplar model there is more accurate judgments for Old exemplars: these judgments benefit from retrieval of identical exemplars with the correct criterion.

One way to predict the relative importance of mental cue abstraction and exemplar memory in the binary and the continuous tasks is by computational considerations (see Juslin et al., 2002). For judgments of a continuous variable—assuming a linear additive model, as people tend to (e.g., Brehmer, 1994)—observation of five exemplars with their criteria is, in principle, sufficient to identify the structure of the task. This system of five linear equations has the unique solution provided by Eq. 1. Given a psychological bias towards linear additive models, the task thus has a well-defined rule-based solution that can be induced from a small number of observations. Binary judgment affords no unique solution, even if the correct function form is assumed and all 16 exemplars are considered. Given the difficulty of inducing a rule-based solution, the participants may have little alternative but to rely on exemplar memory (see Smith & Minda, 2000, for similar arguments).

Note the alternative hypothesis suggested by a *single-systems account* (Nosofsky & Johansen, 2000): that the participants rely on exemplar memory in both tasks. On computational grounds there seems to be no reason why exemplar memory should not be equally applied in both tasks. Both tasks allow a linear combination of criteria stored with exemplars. The hypothesis proposed here is based on a *dual-process account*. Because rule-based knowledge affords better communication and system-

atic elaboration, we expect explicit rule-based processes to be applied when the task structure and the feedback allow participants to induce the task structure, whereas exemplar memory provides a general and flexible back-up system when the task structure or feedback is poor.

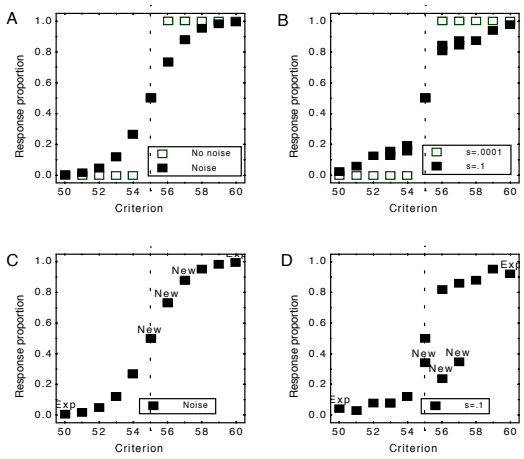

Figure 2: Predictions for the binary task. Panel A: Cue abstraction models with no noise and noise for the complete training set. Panel B: Exemplar model with all similarity parameter *s* equal to .0001 and .1 for the complete set. Panel C: Cue abstraction model with noise for the constrained set. Panel D: Exemplar model with similarity parameter *s*=.1 for the constrained set.

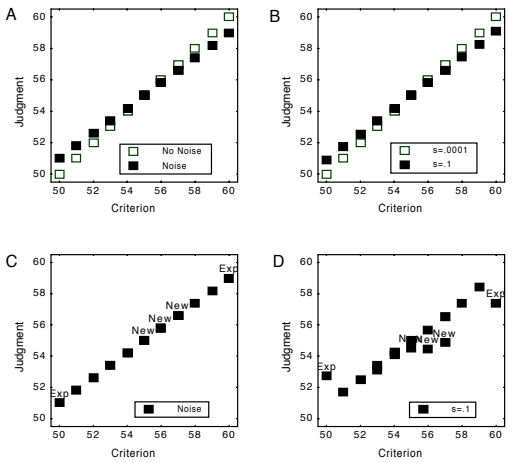

Figure 3: Predictions for the continuous task. Panel A: Cue abstraction models with no noise and noise for the complete training set. Panel B: Exemplar model with all similarity parameter *s* equal to .0001 and .1 for the complete set. Panel C: Cue abstraction model with noise for the constrained set. Panel D: Exemplar model with similarity parameter *s*=.1 for the constrained set.

Method

Participants

Sixty-four persons participated in the experiment (35 women and 29 men, with an average age of 23.5 years). All participants were undergraduate students at Umeå University and rewarded with 70 SEK (app. 7 US $) for their participation in the experiment.

Materials and Procedure

The written instructions informed the participants that there were different subspecies of a Death bug. The subspecies differed in toxicity between 50 and 60 ppm, toxicity below 55 is harmless and toxicity above 55 is dangerous. In the binary task condition, the instruction asked the participants to categorize the subspecies into dangerous and harmless. The training phase provided trial-by-trial outcome feedback about the binary criterion ("This bug is dangerous"). In the continuous task condition, the task was to directly estimate the *toxicity* of the subspecies as a number between 50 and 60. In training, the participants received feedback about the continuous criterion ("This bug has toxicity 57 ppm"). The question on the computer screen was "Is this subspecies harmless or dangerous? (binary task)" or "What is the toxicity of this subspecies? (continuous task)".

The subspecies varied in terms of four binary cues; leg length (short or long), nose length (short or long), spots or no spots on the fore back, and two patterns on the buttock. The cues had the weights 4, 3, 2, and 1 (Eq. 1). The weights determine the portion of toxicity that each cues adds to the total amount. In the analogue stimulus condition, the participants were presented with pictures of the subspecies, in the propositional stimulus condition they were presented with four propositions that provided information about the cue values.

The training phase consisted of 220 trials, where the 11 training exemplars in Table 1 were presented 20 times each. The remaining five exemplars were omitted in the training phase. Two different training sets were used (Sets A and B in Table 1). In Set A, Exemplars 5, 6, and 7 were omitted; in Set B, Exemplars 4, 9, and 10. The exemplars in the two training sets were pair-wise equal in toxicity and the omission of these exemplars was thus counterbalanced across the training sets.

In the test phase, all participants judged all 16 exemplars, twice with an analogue stimulus format and twice with a propositional stimulus format. The stimulus formats were presented in two 2x16 blocks, the order of which was counterbalanced across the participants. No feedback was provided in the test phase. Half of the participants were trained with analogue stimuli and the other half with propositional stimuli, whereas all participants were tested with both presentation formats.

Results

Results and model fits were collapsed over the analogue and propositional conditions, as the aim of this paper is to investigate the relative importance of mental cue abstraction and exemplar memory in binary and continuous tasks. Figure 4 presents model fits (r^2 & *Root Means Square Deviation*) and mean judgments.

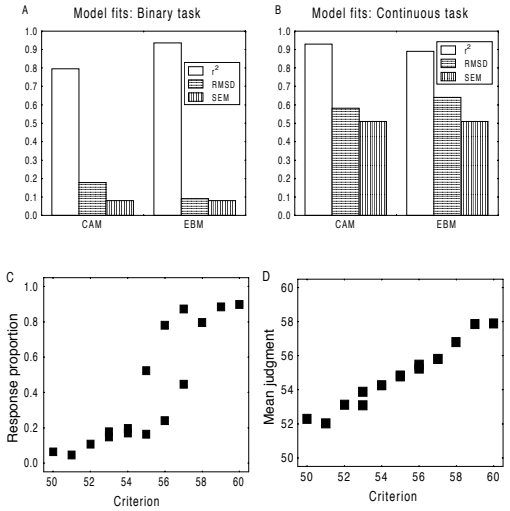

Figure 4: Panel A: Model fits for the binary task. Panel B: Model fits for the continuous task. Panel C: Response proportions in the binary task. Panel D: Mean judgments in the continuous task.

Inspection of Figure 4C suggests exemplar effects in the binary judgment task (notice the difference between new and old exemplars with criterion 55, 56, and 57). The mean difference in proportion of dangerous decisions between old and new intermediate exemplars was -.33 (95% CI: -.45 — -.21). Both the cue abstraction model and the exemplar model were fitted to data. The four parameters in each model were estimated with a Quasi-Newton procedure that minimized the sum of squared deviations between data and model predictions for the last 110 trials in the *training block*. These parameters were used to predict data in the *test phase* (i.e., all free parameters were determined by training data and thus produce cross-validation for training exemplars and genuine predictions for new exemplars).

The exemplar model is clearly superior in the binary judgment task. The model fit indexes in Figure 4A for the binary task, *RMSD* and r^2, suggests predominant use of exemplar processes with model fits almost identical to the mean standard error in data (*SEM*) and r^2 above .90. In the continuous task, the model fits are more ambiguous, although there is a slight advantage for the cue abstraction model (Figure 4B). However, there are nonetheless signs of exemplar effects (e.g., the judgments for the extreme exemplars are at the level of, or below, the judgments for the second-to-most extreme). A comparison between the two tasks revealed that the percentage of participants showing exemplar effects dropped from 81% in the binary task to 63% in the continuous task (p = .06).

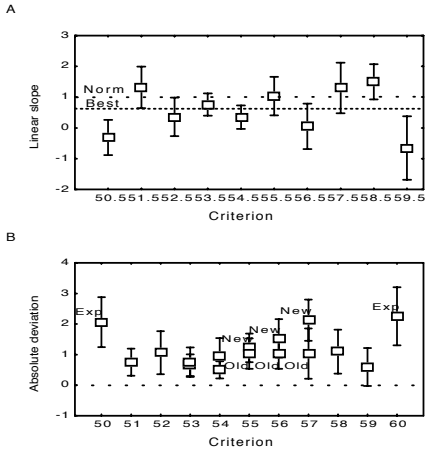

Figure 5: Exemplar effect in the data for continuous judgments with the same training and test stimuli. Panel A: Mean difference (slope) with 95% CI between each successive data point. Panel B: Mean absolute difference from the correct value for each data point.

The signs of exemplar effects are evident in Figure 5 presenting data for continuous judgments with the same training and test stimuli. Panel 5A plots the difference between each successive data point (slope) in a graph like Figure 4D. This slope is 1 for perfect judgments. Panel 5A also provides the slope of the best fitting linear regression of the mean judgments on the criterion. Panel 5B presents the mean absolute error of judgment for each criterion. It is clear that the slopes turn negative for the extreme criteria (inability to extrapolate) with more error in the judgments for new exemplars. There is poorer ability to extrapolate the continuous judgments when training and test stimuli were in the same format ($F(1, 30) = 4.62$, $p = .04$), thus suggesting more exemplar retrieval.

The ambiguous results for the continuous judgments suggest that the group-level data may actually be a mix of the two processes. Investigation of individual participants indeed revealed individual differences. Some participants relied on cue-abstraction, others on exemplar retrieval (Figure 6). Somewhat arbitrarily, but as bench-mark, we deemed best-fitting models accounting for more than 70% of the variance in individual data as producing acceptable fit. On this criterion, 11 participants (34%) were best accounted for by the exemplar model, 13 (41%) were best accounted for by the cue abstraction model, whereas 8 (25%) were not accounted for (r^2<.7 for both models). In sum: although there are exemplar effects also with a continuous criterion, there

is an increased prevalence of cue abstraction with some participants clearly relying on cue abstraction.

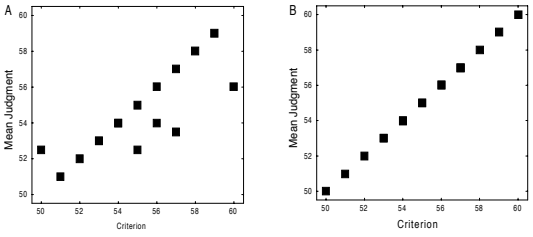

Figure 6: Individual-participant data: Panel A exemplifies a participant guided by exemplar retrieval and panel B a participant guided by cue abstraction.

Discussion

The question addressed in this article is why the theoretical conclusions from categorization and multiple cue judgment research are different, considering that the task structure is so similar (Figure 1). Perhaps, the most salient difference between the paradigms is that categorization often involves a binary whereas multiple-cue judgment often involves a continuous criterion.

The results suggest that the differential emphasis in the conclusions is not accidental to different research traditions, with more cognitive modeling in categorization research and more statistical modeling in multiple cue judgment research. Changing the criterion from binary to continuous thus creates a shift from exemplar memory to a mix of exemplar- and rule-based processing that involves cue abstraction in training and cue-integration at the time of judgment. In the continuous judgment condition just as many individual participants extrapolated appropriately and relied on cue abstraction as on exemplar memory. Figure 6 highlights the individual differences in preferred representational mode (see Shanks & Darby, 1998, for similar results). These results raise the question of the appropriateness of the routine procedure of applying quantitative models to group-level data. The exemplar retrieval with continuous judgments moreover seems to increase when training and test conditions coincide.

There is no reason why exemplar memory should not be used in both tasks (as it indeed was by some participants). Exemplar retrieval is an equally efficient way to solve both tasks. However, it seems that as soon as the feedback is informative enough, people eagerly induce explicit rule-based representations, corresponding to the "rule-bias" suggested by Ashby et al. (1998). This suggests that people change between qualitatively distinct representation levels depending on the task properties (Ashby et al., 1998; Jones et al, 2000; Juslin et al., 2002). Jones et al. showed that people spontaneously tend to integrate cues in a task like the one used here, either explicitly by cue abstraction or implicitly by exemplar retrieval. A principled understanding of the interplay between – and properties of – these distinct levels of representation in human judgment and categorization should be a prime goal of cognitive science.

Acknowledgments

Bank of Sweden Tercentenary Foundation supported this research.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105,* 442-481.

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

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

Einhorn, J. H., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Regression models and process tracing analysis. *Psychological Review, 86*, 465-485.

Jones, S., Juslin, P., Olsson, H., & Winman, A. (2000). Algorithm, heuristic or exemplar: Processes and representation in multiple-cue judgment. In L. Gleitman, & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 244-249). Hillsdale, NJ: Erlbaum.

Juslin, P., Olsson, H., & Olsson, A-C. (2002). *Abstract and concrete knowledge in categorization and multiple-cue judgment*. Manuscript submitted for publication. Department. of Psychology, Umeå University, Umeå, Sweden.

Juslin, P., & Persson, M. (2000). *PROBabilities from EXemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge*. Manuscript submitted for publication. Department. of Psychology, Umeå University, Umeå, Sweden.

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

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

Shanks, D. R., & Darby, R. J. (1998). Feature- and rule-based generalization in categorization. *Journal of Experimental Psychology: Animal Behavior Processes, 24,* 405-415.

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

Smith, E. R., & Zarate, M. A. (1992). Exemplar model of social judgment. *Psychological Review, 99,* 3-21.