Journal of Experimental Psychology:
Learning, Memory, and Cognition
2006, Vol. 32, No. 1, 163–179

0278-7393/06/$12.00 DOI: 10.1037/0278-7393.32.1.163

On the Role of Causal Intervention in Multiple-Cue Judgment: Positive and Negative Effects on Learning

Tommy Enkvist
Uppsala University

Ben Newell
University of New South Wales

Peter Juslin and Henrik Olsson
Uppsala University

Previous studies have suggested better learning when people actively intervene rather than when they passively observe the stimuli in a judgment task. In 4 experiments, the authors investigated the hypothesis that this improvement is associated with a shift from exemplar memory to cue abstraction. In a multiple-cue judgment task with continuous cues, the data replicated the improvement with intervention and participants who experimented more actively produced more accurate judgments. In a multiple-cue judgment task with binary cues, intervention produced poorer accuracy and participants who experimented more actively produced poorer judgments. These results provide no support for a representational shift but suggest that the improvement with active intervention may be limited to certain tasks and environments.

Keywords: intervention, observation, learning, multiple cue judgment

Information about our environment can be acquired in a variety of ways. We learn through instruction from others, by passive observation, and by acting on our environment and observing the consequences of our interventions. These methods of acquiring information are all fundamental to our ability to adapt and function successfully in the environment. For example, when searching for Karl–Johan mushrooms in a Swedish forest in autumn, we might consult a reference book to find a picture of the mushroom (information via instruction) then compare the picture with mushrooms we spot in the forest (information via observation)—to ensure we picked the nontoxic variety—and then, in our attempts to produce a delicious soup, we might experiment with the ingredients (information via intervention).

Despite the prevalence of all three of these strategies for learning about our environment, there has been, up until relatively recently, an imbalance within the fields of judgment and categorization in the focus of research into the different strategies. Some research has examined the role of instruction in learning multiple-cue tasks (J. R. Anderson & Fincham, 1994; Johansson & Brehmer, 1979; Kuylenstierna, 1998), and numerous experiments address how learning through observation results in the ability to categorize stimuli (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Juslin, Jones, Olsson, & Winman, 2003; Juslin, Olsson, & Olsson, 2003; Medin & Schaffer, 1978; Nosofsky & Johansen, 2000; Nosofsky & Palmeri, 1997; Nosofsky, Palmeri, & McKinley, 1994; Smith & Minda, 2000) or to make predictive judgments (Brehmer, 1994; Cooksey, 1996; Einhorn, Kleinmuntz, & Kleinmuntz, 1979; Hammond & Steward, 2001). There is, however, relatively little work examining the role that intervention might play in learning in multiple-cue tasks and the subsequent effect it might have on the nature of the cognitive representations.

The research on learning from intervention in causal reasoning tasks suggests that, perhaps not surprisingly, learning is promoted by the possibility to intervene with the system under study (Lagnado & Sloman, 2004). In this article we extend this finding by investigating the role of intervention in a multiple-cue judgment task where the participants use a number of cues to infer a criterion. Research with this task (Juslin, Jones et al., 2003; Juslin, Olsson, & Olsson, 2003) has identified two qualitatively different cognitive processes that can underlie the performance of 'observers'. The first, inspired and motivated by research on categorization, emphasizes exemplar memory and assumes that people make judgments by retrieving similar exemplars from memory (Medin & Schaffer, 1978; Nosofsky & Johansen, 2000). The second, derived from research on multiple-cue judgment, stresses the controlled integration of explicit knowledge of cue-criterion relations abstracted in training (Einhorn et al., 1979).

In this context we investigate the relationship between active intervention and the cognitive representations acquired and we argue that the benefit of intervention should be especially large if

Tommy Enkvist, Peter Juslin, and Henrik Olsson, Department of Psychology, Uppsala University, Uppsala, Sweden; Ben Newell, School of Psychology, University of New South Wales, Sydney, Australia.

This research was supported by the Bank of Sweden Tercentenary Foundation, the United Kingdom Economic and Social Sciences Research Council, and the Australian Research Council. Portions of this work were conducted while Ben Newell was a visiting research fellow at Uppsala University, and he would like to thank the Department of Psychology for their generosity and hospitality. We also thank Patrik Hansson, Håkan Nilsson, Anna-Carin Olsson, and Anders Winman for comments on previous versions of this article, and David Lagnado for many enlightening discussions about the nature of intervention.

Correspondence concerning this article should be addressed to Tommy Enkvist, Department of Psychology, Uppsala University, Uppsala, SE 751 42 Sweden. E-mail: tommy.enkvist@psyk.uu.se

one engages in cue abstraction. Accordingly, one hypothesis in regard to why intervention affords a benefit over observation is that it promotes knowledge representation in the form of abstract cue-criterion relations rather than memory for exemplars (Juslin, Jones, et al., 2003; Juslin, Olsson, & Olsson, 2003). Because causal intervention with the environment allows controlled "experimentation," for example, by keeping all cues but one constant to investigate its effect on the criterion, arguably it should become easier to abstract the relations between individual cues and the criterion. One prerequisite for this shift toward cue abstraction is that people spontaneously engage in this sort of "experimentation."

Our approach in this article is to take a task for which we have prior evidence concerning the nature of learning and the acquired representations and investigate the effect of contrasting two strategies—observation and intervention—for acquiring information. The task involves a probe described by four cues that have either continuous (Experiment 1) or binary values (Experiments 2, 3, and 4). In the binary case, the probe is a fictitious "death bug" described by cues, such as "short legs" or "brown body," the combination of which determines the "toxicity" of the bug, and in the continuous case each cue is described by a numeric value, such as "leg length = 5′" or "back color = 6". The participant's task is to learn to judge the toxicity of each bug. One group—the observers—learn through passive exposure to bugs with different combinations of features; a second group—the interveners—learn by actively constructing bugs from constituent features to achieve a given level of toxicity.

Intervention, Observation, and Learning

Other than a general intuition that intervention might lead people to think more analytically, and test their own hypotheses about how combinations of features determine the criterion in our experimental task, is their any empirical evidence indicating the role intervention might play in such a task? In the Western scientific tradition, the advantage of experimentation over simple observation has long been recognized (Mill, 2002), but as noted above, in the areas of categorization and multiple-cue judgment, research on intervention is scarce. Consideration of recent advances in understanding how people learn causal structure, however, leads us to be optimistic about the advantages intervention might confer.

Klayman (1988) used a task in which there was no salient causal structure to infer but that did necessitate the discovery of predictive cues. The participants were presented with a computer-controlled graphic display in which geometric figures appeared in various locations. On each trial a figure appeared that could be one of three shapes and shadings. An asterisk then appeared on the screen and a straight line or trace was drawn out from that asterisk. The task was to learn to predict whether a particular trace would stop before it reached the edge of the display or whether it would simply go off the screen: and, if it were to stop, would it do so where they thought it would. Klayman's experiments showed that participants who were free to design their own screens and locate the shapes and trace origins anywhere on the display did better than participants who just observed a random selection of trials. Within the intervention group there was wide individual variability in the degree and quality of experimentation engaged in, and there was a strong association between the quality of experimentation and the success in discovering predictive cues. "Good" experimenters only changed one variable between trials when testing a hypothesis and achieved more accurate predictions than the "bad" experimenters who changed several variables between consecutive trials.

In categorization research there is an increasing interest for the activity that people engage in when learning about categories (Ross, 2000; Ross & Warren, 2002). For example, learning about categories by inferring unknown properties of the category members rather than by classifying them into categories is one step from passive observation toward a more active learning context. By actively choosing the values of missing features inference learning promotes comparison within a single category and the underlying commonalities are more likely to be detected (Markman & Ross, 2003). Interaction, moreover, forces the learner to focus both on the observable features and on the relationships between features that are less easily observed but important for how the instance is used (Ross, 1996).

Advances in the formal modeling of causal relations (Pearl, 2000) has stimulated renewed interest for casual reasoning and its role in learning (Gopnik et al., 2004; Rehder, 2003; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). Lagnado and Sloman (2004), for example, used a trial by trial based learning paradigm in which participants obtained probabilistic data about causally related events either through observing sequences (e.g., seeing a high fuel temperature and a low combustion-chamber pressure leading to the launch of a rocket) or through intervention (e.g., setting both temperature and pressure to either high or low and then observing whether a rocket launched or failed). The results showed a clear advantage for interveners in terms of their ability to subsequently select the causal model likely to have generated the data (from an array of possible models).

Other studies (e.g., Steyvers et al., 2003) have also shown the power of intervention in causal learning by demonstrating that learning that links observation and intervention guides intervention toward maximally informative targets. In a study by Sobel (2004) the participants in an intervention condition performed much better than participants in an observation condition (66% correct compared with 35% correct) when learning probabilistic causal structures. Indeed, preschool children already seem highly sensitive to quite subtle causal cues in their attempts to understand their environment (Gopnik et al., 2004).

Both intuition and data therefore suggest fundamental differences between the performance of observers and interveners. There is, however, less research that has charted the limiting conditions of the improved learning with intervention, or that has addressed more specific hypotheses about the reasons for the benefit. The experiments reported below therefore investigate if the benefit is mediated by qualitative differences in the cognitive representations. We use an experimental procedure designed to distinguish between two candidate representations: exemplars and abstracted cue-criterion relations.

Judgment Task and Cognitive Models

The distinction between exemplar memory and cue abstraction emphasized in Juslin, Olsson, & Olsson (2003) is paralleled by a number of approaches to cognition in which comparisons are made between explicit rule-based, analytical knowledge on the one hand and similarity-based knowledge on the other (Ashby et al., 1998;

Erickson & Kruschke, 1998; Hammond, 1996; Sloman, 1996; E. E. Smith, Patalano, & Jonides, 1998). One idea is that people prioritize explicit rule-based processes (Ashby et al., 1998) but fall back on exemplar memory when the explicit system fails—perhaps because the cue-criterion relations are not readily abstractable (Juslin, Jones et al., 2003; Juslin, Olsson et al., 2003).

We rely on variations of an experimental paradigm designed to help distinguish between cue abstraction and exemplar memory in a multiple-cue judgment task (Juslin, Jones, et al., 2003). The task originally involves a probe defined by four binary cues and requires judgment of a continuous criterion. Judgments are made in a training phase in which feedback about the correct criterion is provided after every judgment. The cover story involves the toxicity of subspecies of the exotic (but fictitious) death bug. In Experiment 1, the task has the same cover story and a probe with four cues, but it involves judgment based on four continuous cues. The cues C_1, C_2, C_3 and C_4 take on 11 discrete values between 0 and 10 and the toxicity c of a subspecies is a linear additive function of the cues:

$$c = 500 + 4 \cdot C1 + 3 \cdot C2 + 2 \cdot C3 + 1 \cdot C4. \quad (1)$$

The Criterion c is computed by assigning cue one, C_1, the largest weight and cue number four, C_4, the least importance. In Experiment 1, two cues are positively linearly related and two cues are negatively linearly related to the criterion. The task is summarized in Table A1 of Appendix A.

Cue Abstraction

One way to address this task is by inferring the relationship between each individual cue and the criterion at training, forming abstract representations of the cue-criterion relations that are integrated at the time of judgment (Einhorn et al., 1979). Juslin, Karlsson, and Olsson (2004) derived the implications of regarding the judgment process as a controlled and capacity constrained process of sequential adjustment of an estimate of the criterion. In regard to cue abstraction two implications are relevant here. First, because of constrained working memory capacity that only allows comparison of pairs of criterion values at any moment the cue criterion relations tend to take the form of *linear slopes* (differences; i.e., detection of nonlinear relations requires consideration of at least three criteria).

Second; it is only possible to ascertain the relationship between an individual cue and the criterion when the effects of the other cues have been controlled for. If someone, for example, encounter two exemplars in rapid succession, one with cue values 1, 1, 1, 1 and Criterion 60, the other with cue values 0, 1, 1, 1 and Criterion 56, he or she might well infer that the difference on the first cue explains the difference of four units in the criterion. By contrast, it is difficult to ascertain the effect of the first cue if multiple cues differ between the exemplars. Cue abstraction therefore tends to be highly dependent on making controlled observations. In essence, because of the constraints on controlled judgment the knowledge of cue-criterion relations become linear, orthogonal caricatures of the relations in the environment.

At the time of judgment the process of successive sequential adjustment is constrained to add up the impact of individual cues (H. Anderson, 1981; Hogarth & Einhorn, 1992). This view of the cue abstraction process suggests that people should have explicit knowledge of the cue-criterion relations (but not necessarily of the additive cue integration rule that emerges from the sequential process), that this knowledge should take a very specific form, and that judgment by cue abstraction should only be efficient in environments where the criterion is well approximated by linear additive combination of the cues (Juslin et al., 2004). A quantitative implementation of cue abstraction is provided in Appendix B. Of relevance here, this view suggests that the possibility to conduct controlled experimentation in the intervention condition, for example, by systematically creating successive exemplars that differ with respect to a single cue, should be crucially important for cue abstraction.

Exemplar Memory

When relying on exemplar memory, the judgment is based on retrieving similar previous exemplars from long term memory and using the criteria of these retrieved exemplars to infer the criterion of the new exemplar (e.g., Medin & Schaffer, 1978; Nosofsky & Johansen, 2000). The exemplars stored in memory are activated as a function of their overall similarity to the new exemplar and, therefore, the process need not involve explicitly abstracted knowledge of how individual cues relate to the criterion.

In contrast to cue abstraction, exemplar memory is not premised on very specific assumptions about the cue-criterion relations in the task environment (e.g., that they involve linear additive cue combination) and thus allows accurate performance also in, for example, tasks that involve multiplicative cue combination (Juslin et al., 2004). Also in contrast to cue abstraction, the ability to store and retrieve similar exemplars should benefit much less from the opportunity to conduct controlled experimentation. Indeed, to the extent that the task is most efficiently addressed with exemplar memory, intervention, that might encourage cue abstraction, could actually distract the participants from the relatively more efficient exemplar memory process. A quantitative exemplar model is provided in Appendix B.

Although, behaviorally, cue abstraction and exemplar memory mimic each other perfectly in many situations (Juslin, Olsson et al., 2003) one key to distinguish between them is by withholding in training some exemplars that require extrapolation and interpolation. The ability to extrapolate and to interpolate measures how well the cue-criterion relations have been abstracted (DeLosh, Busemeyer, & McDaniel, 1997). If participants make correct judgments for the extreme exemplars presented only in the test phase (i.e., extrapolate) it is likely that the participant has figured out the cue-criterion relations, which is suggestive of cue abstraction. If exemplar memory is used, the participant is unable to extrapolate beyond the range of stimuli seen in training because the judgment is a weighted average of the criteria in training. With cue abstraction, when judging new exemplars in the middle range, there will be no systematic differences between new and old exemplars. With the exemplar model, on the other hand, old exemplars are judged correctly more often than are new exemplars because only old exemplars with correct criteria can be retrieved from memory. Figure 1 illustrates the predictions both for continuous cues, as in Experiment 1 below, and for binary cues, as in Experiments 2, 3, and 4 below (see Appendix B for details on the models). In general, cue abstraction affords accurate judgment for both old and new exemplars, whereas exemplar memory implies large errors for

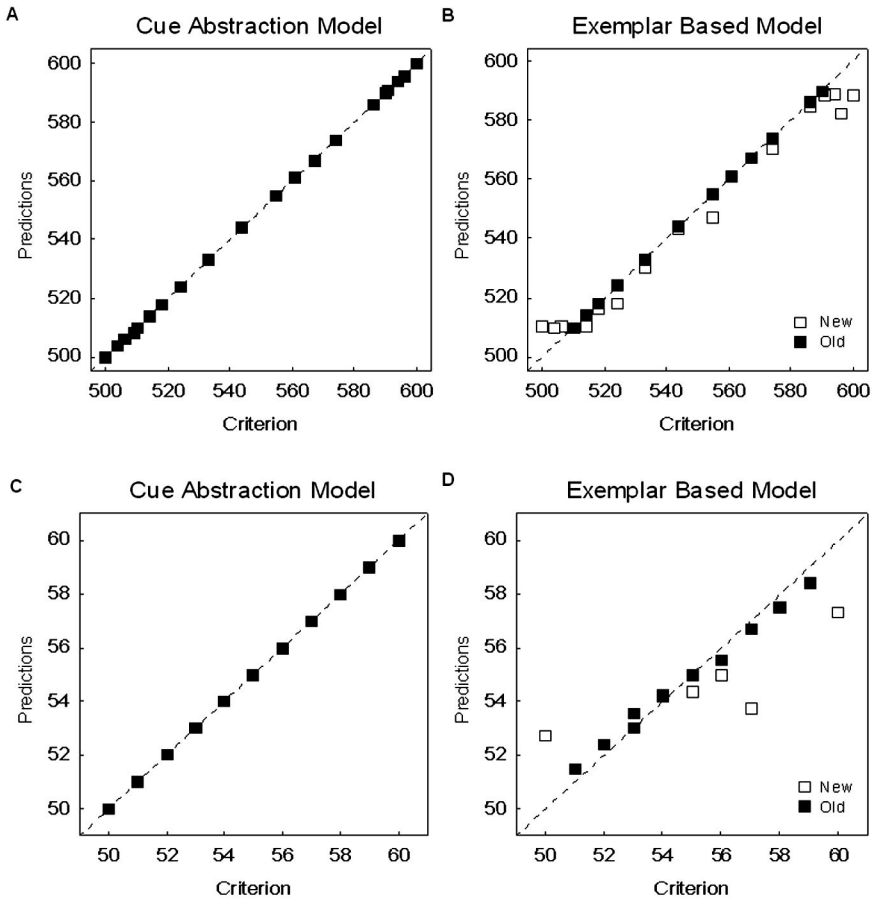

Figure 1. Illustration of typical predictions by the cue-abstraction and the exemplar memory models (see Appendix B for computational details). A: The cue-abstraction model with the correct cue weights in a task with continuous cues. B: The exemplar model with parameters .25 and $c = 10$ in a task with continuous cues. C: The cue-abstraction model with the correct cue weights in a task with binary cues. D: The exemplar model with parameters .25 and $c = 10$ in a task with binary cues. Solid squares represent old exemplars and open squares represent new exemplars.

new exemplars with an inability to extrapolate (Juslin, Olsson, & Olsson, 2003).

In regard to active and causal intervention with the stimuli during training, the framework provided by these models raises a number of more specific questions: (a) Is the beneficial effect of intervention limited to tasks spontaneously addressed by cue abstraction, as suggested by the characterization of these processes provided above (Juslin et al., 2004)? (b) Do people spontaneously engage in controlled experimentation when possible? (c) Several studies (Juslin, Jones et al., 2003; Juslin, Olsson, & Olsson, 2003) suggest that people adapt their processes to the demands imposed by the task. Do we observe a similar shift from exemplar memory to cue abstraction when cue abstraction is facilitated by intervention?

Experiment 1: Continuous Cues

In Experiment 1, we compared the performance by participants who observed (the observation condition) or actively constructed (the intervention condition) the stimuli in the training phase. The aims of Experiment 1 are threefold: The first aim is to replicate the benefit of active intervention observed in previous studies with the sort of multiple-cue judgment used in the present study. The second is to investigate whether people spontaneously realize the possibility to conduct controlled experimentation when they can actively intervene with the stimuli in training. The third, provided that the beneficial effect of intervention is replicated, is to determine whether this improvement is associated with a shift from exemplar memory to cue abstraction.

Interveners are given a target criterion and then asked to construct an exemplar with this criterion by selecting individual cue values. This gives interveners the opportunity to conduct controlled experimentation such as structuring successive created exemplars that differ with respect to only a single cue. This experimentation should promote cue abstraction. In contrast, observers with no control over the cues should find cue abstraction harder; in particular, if it is constrained to, or heavily biased toward, estimating linear slopes between successive exemplars differing with respect to a single cue (Juslin et al., 2004). We hypothesized that

interveners should exploit the possibility to conduct controlled experimentation by more often creating successive exemplars differing with respect to only one cue as compared with the baseline provided by the observers and that, relative to the observers, there should be a shift from exemplar memory to cue abstraction among the interveners.

Method

Participants. Thirty-two undergraduate students (21 women, 11 men) from Uppsala University volunteered. All received payment of approximately 80 SKr (US$10) or course credit. The mean age of the participants was 24.4 years (SD = 2.98; range: 20–32). Participants were randomly divided into two groups, observation and intervention. All participants were tested individually.

Materials and procedure. Fictitious death bugs with continuous cues and a continuous criterion were used as stimuli. Each cue could take on 11 different values, represented in the experiment by a number ranging from 0 to 10. Assigning four cues 11 different values yields 11^4 possible combinations. The criterion range between 500 and 600 is computed from Equation B1 (see Appendix B). Two cues were positively linearly related and two cues were negatively linearly related to the criterion. For the positively related cues toxicity increased when the cue value increased and for the negatively related cues toxicity increased when cue value decreased. The cue order and the cue directions (positive or negative relation) were randomized for each participant.

Experiment 1 contained two phases, a training phase of 120 trials and a test phase of 60 trials. For the observers, each learning trial consisted of the presentation of text-based descriptions (see Juslin, Olsson et al., 2003) of a fictitious death bug species with four continuous attributes (leg length, back color, nose length, and color pattern on neck). Each cue was presented by its text label and a number representing the value of the cue (e.g., leg length = 8, back color = 5, and so on).[1] In the training phase, 40 unique exemplars in the criterion range 510–590 were presented three times and the participants had to judge the toxicity of the presented exemplars. Participants answered the question *What is the toxicity of this bug?* and made a numerical response (the amount of toxic substance measured in ppm) on each trial. Feedback showing the correct toxicity followed the response and remained on the screen until the participant clicked to advance to the next trial. In the test phase, 30 unique exemplars with criterion from 500 to 600 were presented twice. The test phase consisted of 12 old exemplars previously encountered in training and 18 new exemplars, 8 in the extrapolation region and 10 in the interpolation region of the criterion range. (Table A1 of Appendix A contains all exemplars used in the experiment.) As in the training phase, participants had to judge the toxicity of the presented exemplars, but with the feedback omitted.

The interveners saw the same screen layout in training, but instead of judging the toxicity of the bug they had to decide the cue value of each cue in order to create a bug of a given toxicity. The cue value was a number between 0 and 10 entered by the participant. On a trial they were asked, for example, to *Create a bug that has a toxicity of 570 ppm*. After selecting the value of all four features and clicking *create*, feedback on the toxicity of the created bug was given. The 18 new exemplars used in the test phase were not possible to create under training, and an error message occurred every time a participant tried to create one of these exemplars. The test phase was the same as for the observers.

Dependent measure. The experimentation index measures the degree to which participants attempted to conduct controlled experimentation. The experimentation index that ranges between 0 and 4 is the number of cues that are kept constant across two successive exemplars. In the observation condition the experimentation index is determined entirely by the random sequence of exemplars presented and this score thus serves as a random baseline (the expected cue change if exemplars are observed randomly). In the intervention condition the experimentation index is under control of the participants and is a function of their attempts to conduct controlled experimentation. The performance measure that addresses accuracy is root mean square error (RMSE) of the judgment made in the test phase (mean square deviation between judgment and criteria).

Interpolation is measured by the signed difference between the absolute deviations between judgment and criteria for old and new exemplars. If the absolute deviations from the criterion are equally large this difference is 0, suggestive of cue abstraction. If the deviations are larger for new interpolation exemplars, as predicted by the exemplar model, the difference is negative. Extrapolation is measured by the signed deviation from the prediction by a linear regression model with judgment as dependent variable and criterion as the independent variable. If the judgments for the extreme exemplars are as extreme as expected from linear extrapolation from the training exemplars the deviation is 0. If participants fail to extrapolate, as predicted by the exemplar model (Figure 1B), it is negative. The representation index (RI) is the sum of the interpolation and extrapolation measures and a RI of 0 is suggestive of cue abstraction, a significantly negative RI is suggestive of exemplar memory.

Model fit is analyzed by using data from the training phase in each task to create predictions by the exemplar based model (EBM) and the cue abstraction model (CAM; see Appendix B). For each participant the four best-fitting parameters for each model were ascertained by minimizing the root mean square deviation (RMSD) between model predictions and the last judgment made for each of the 40 exemplars in the training phase (see Table A1). These parameters were then used to predict how participants should perform in the test phase with both old and new exemplars. Model fit is measured by the coefficient of determination (r^2) and the RMSD between predictions and test-phase data. The cue abstraction model allows analytic derivation of the best-fitting parameters. Parameters for the exemplar model were estimated by the Quasi-Newton method in the MathCAD software.

Results

Performance. Experimentation indices were computed for each block of 20 trials in the training phase, producing six indices for each participant. This index was entered as the dependent variable in an analysis of variance (ANOVA) with training condition (observation vs. intervention) as a between-subjects independent variable and training block (Blocks 1 to 6) as a within-subjects independent variable. The ANOVA yielded a significant main effect of condition, $F(1, 180) = 102.39$, $MSE = .111$, $p = .000$, a significant main effect of block, $F(5, 180) = 4.14$, $MSE = .111$, $p = .001$, and a significant interaction, $F(5, 180) = 3.43$, $MSE = .111$, $p = .006$. As illustrated in Figure 2, the key effect is the interaction: In the observation condition, the experimentation index is flat at its random base-level throughout training; for the intervention condition it is clearly higher early in training, and later levels off as the participants learn the task.

A one-way ANOVA with RMSE from the test phase as dependent variable and intervention versus observation training as independent variable shows a significant difference in performance between the groups in favor of intervention $F(1, 30) = 9.9$, $MSE =$

[1] Presenting the stimuli as visual objects (e.g., actual bugs) with actual visually presented dimensional values versus stating the cue values verbally in written text could in principle affect the cognitive processes involved. With the present experimental paradigm, however, we have been repeatedly unable to find any differences in the results (performance or model fit) between presenting stimuli in a text format or as visual objects (see, e.g., Juslin, Jones, et al., 2003; Juslin, Olsson, & Olsson, 2003; Nilsson, Olsson, & Juslin, 2005).

75.89, $p = .004$ (see Table 1). Also in the last block of training, intervention has a significantly lower RMSE indicating improved learning, $F(1, 30) = 6.36$, $MSE = 19.18$, $p = .017$. Figures 3A and 3B, which present the mean judgments, clearly illustrate the superior judgments with intervention. The correlation between each intervener's experimentation index computed across the first 40 trials in training and his or her judgment accuracy at test as measured by RMSE ($r_{32} = -.48$, $p = .005$) indicates that more experimentation did improve judgment accuracy.[2] In training, the observers saw 40 unique exemplars presented three times. The mean number of exemplars created in the intervention condition was 50.69.

Representation. There were no significant differences in the RI between the conditions, $F(1, 30) = 1.4$, $MSE = 75.62$, $p = .25$, but the mean RI was negative (−2.68) for the observers and positive (.96) for the interveners. This may suggest that the interveners were able to extrapolate and interpolate more effectively than the observers but that the RI is not significantly separated from 0 indicates large individual differences. The correlations between the experimentation index and the RI were low and not significant.

Model fit. The exemplar and cue abstraction models (see Appendix B) were fitted separately to the training data for each participant and then applied to predict the judgments in the test phase. As presented in Table 1, the fit was very similar for both models. Accordingly, entering the RMSD between model predictions and data as dependent variable in a split-plot ANOVA with training condition (observation vs. intervention) as the between-subjects variable and model (cue abstraction vs. exemplar memory) as the within-subjects variable yielded no significant main effect of model, $F(1, 30) = .63$, $MSE = 144.73$, $p = .434$, no significant main effect of training condition, $F(1, 30) = .07$, $MSE = 23.41$, $p = .788$, and no significant interaction, $F(1, 30) = .01$, $MSE = 23.41$, $p = .914$. Further analysis revealed that the predictions by the two models for these data proved to be extremely highly correlated. For example, if the two models are fitted to the group-level data and predictions are computed for the items in the test phase, the correlation between the predictions exceeds .99.

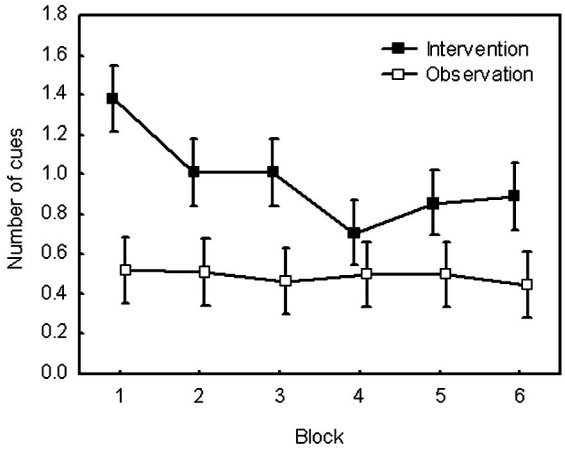

Figure 2. The mean experimentation index for the observation and intervention conditions of Experiment 1 plotted as a function of training block, where each block consists of 20 trails. The means for observation define the change of cues expected by chance.

Table 1
Means for All Dependent Measures in Experiment 1

Measure	Observation	Intervention
RMSE	21.13	11.44
Representation index (RI)	−2.68	0.96
EBM (RMSD)	19.08	16.83
EBM (r^2)	0.52	0.69
CAM (RMSD)	18.88	16.37
CAM (r^2)	0.55	0.67

Note. RMSE = root mean square error; EBM = exemplar-based model; RMSD = root mean square deviation between model predictions and data; r^2 = coefficient of determination, variance accounted for by the model; CAM = cue abstraction model.

Discussion

Experiment 1 replicated the finding from previous studies with other tasks (e.g., Klayman, 1988; Lagnado & Sloman, 2004) that active causal intervention promotes more efficient learning. Not only did the interveners produce much more accurate judgments, accuracy of judgment was correlated with the degree to which the participants spontaneously engaged in controlled experimentation. Although the results suggest that the participants conducted more controlled experimentation in the intervention condition, and benefited from it, that the mean experimentation index in Figure 2 is well below 3 (keeping all but one cue constant)—the most common pattern was to change multiple cues across trials.[3]

[2] Note that the experimentation index is relevant also to the observation condition because if the presentation of training exemplars is randomized, different participants will get a larger or smaller number of successive sequences where all but one cue is kept constant simply as a matter of sampling error. In this case, the negative correlation between the experimentation index and the RMSD shows that those participants who by sampling accident received a larger number of controlled sequences where all but one cue were constant were able to indeed benefit from this and produced more accurate judgments. However, in all three experiments reported in this article the sign and the magnitude of the correlation between the experimentation index and the RMSD are similar when computed across only the interveners that actively controlled the degree of experimentation (i.e., −.38 in Experiment 1 as compared with −.48 when computed across both groups, .17 in Experiment 2 as compared with .19 across both groups, and .57 in Experiment 3 as compared with .50 across both groups).

[3] In later trials of training, the relative lack of experimentation (i.e., keeping all but one cue constant across successive trials) could be explained by the possibility that the participants had already acquired knowledge and therefore always produced the criterion that was asked for. Because the sequence of criteria is the same in the observation and intervention conditions, if participants have perfect knowledge and always produce the criterion asked for, they trivially get the same experimentation index as in the observation condition, completely controlled by the sequence of criteria. This, however, does not apply to the early trials of training, and already in the first block the experimentation index is 1.4, far below what is expected if they experimented on every trial (see Figure 2).

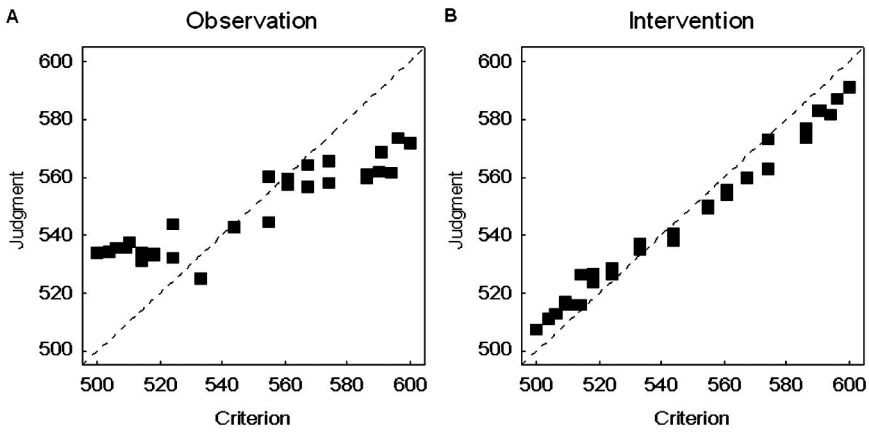

Figure 3. Data from Experiment 1. A: Mean judgments in the observation condition plotted against the criterion. B: Mean judgments in the intervention condition plotted against the criterion.

The hypothesis that the improved performance is mediated by a shift from exemplar memory to more efficient abstraction of the underlying cue-criterion relations was not clearly supported by the data. The data from the intervention group suggests cue abstraction in regard to most dependent measures, low RMSE, nonnegative RI, and a good model fit for the cue abstraction model. This is in line with our hypothesis. But for the observation group we did not get clear support that favored either the cue abstraction model or the exemplar model, although numerically all measures except the model fit suggested a greater prevalence of exemplar memory. Two reasons for the lack of statistical differentiation between the fit of the two models are, first, that for the data obtained in the experiment the models produce predictions that are extremely highly correlated and, second, there is likely to be considerable individual differences in these data (Juslin, Olsson, & Olsson, 2003).

Previous research with this paradigm (Juslin, Jones et al., 2003; Juslin et al., 2004; Juslin, Olsson, & Olsson, 2003) has shown that the task structure itself is an important determinant of the cognitive processes, quite aside from the activity one engages in during training. One possibility is that a task with continuous cues and criterion strongly triggers analytic thinking in both the observation and intervention conditions. On this interpretation the results indicate that in a task spontaneously addressed by cue abstraction there is improvement from active intervention, but because cue abstraction is so prevalent already in the observation condition it is difficult to verify such a shift from exemplar memory to cue abstraction.

Experiment 2: Binary Cues and a Stop Criterion

To further investigate the hypothesis that intervention is linked to a shift in the representation, we turned to a binary multiple-cue judgment task that in previous studies has been more clearly associated with exemplar effects in an observation training regime (Juslin, Olsson, & Olsson, 2003). If intervention triggers more analytic thinking, a task in which exemplar memory is relatively more prevalent in observation training is a better point of departure for detecting the hypothesized shift from exemplar memory to cue abstraction.

In Experiment 2, we used the task with binary cues (Juslin, Olsson, & Olsson, 2003). In this task, each exemplar consists of four binary cues that predict a continuous criterion. This task therefore involves only 16 unique exemplars (see Table A2 of Appendix A). The use of fewer exemplars has been argued to promote exemplar processes (J. D. Smith & Minda, 2000), and earlier experiments with this task suggests that there are clear signs of exemplar memory, although there are also large individual differences (Juslin, Olsson, & Olsson, 2003). One possibility, therefore, is that we should observe the representational shift in this task because the binary cue structure is more conducive to reliance on exemplar memory.

An alternative possibility, arguably more in line with the lack of a statistically significant effect of observation versus intervention on the fit of the two models, is that, again, the task structure itself is such a powerful determinant of the processes that we again fail to observe the representational shift. This hypothesis comes in two varieties. First, it could be the case that the binary task is dominated by cue abstraction. If so, we would expect to replicate the results from Experiment 1 with superior learning for the interveners. Second: it could be that exemplar memory dominates. If the task itself spontaneously and strongly invites exemplar memory, then encouraging cue abstraction by allowing participants to conduct controlled experimentation may have a less positive effect than in Experiment 1, and it could even distract from a more efficient memorization strategy.

In research on categorization learning, Markman and Ross (2003) emphasized the importance of the fit between the training and test conditions, as typically epitomized in the notion of *transfer appropriate processing* (Morris, Bransford, & Franks, 1977). On this view, memory is best when the training and test conditions involve the same cognitive processes. One limitation of Experiment 1 is that whereas the observers performed both training and test judgments in the same observation format, the interveners trained with intervention but were tested with observation. In Experiment 2, we therefore examined the role of training activity (observation vs. intervention) and test activity in a fully factorial design. Will performance improve if interveners have the same context both at training and at test, and will performance deterio-

rate in the observation condition if the test phase changes to intervention?

Experiment 1 also highlights what could be a confounding variable in Experiment 2. In Experiment 1, intervention produced more effective learning, resulting in better performance both at training and at test. Improved learning may as such affect the dominating mode of representation. Specifically, more training and more automaticity in performing a task is often associated with a shift from analytical thinking toward exemplar memory (Logan, 1988; Nosofsky & Palmeri, 1997). Therefore, if intervention promotes learning, and this instills more exemplar memory, the improved learning as such might conceal the presence of a shift from exemplar memory to cue abstraction.

We addressed this potential problem by performing two separate but related experiments: One pair of observation and intervention groups were trained until they reached the same predetermined level of accuracy (Experiment 2); another pair of observation and intervention groups were trained with the same predetermined large number of training trials (Experiment 3). In this way, we were able to assess the representations both when the groups were equated in terms of accuracy and in the number of training trials. In Experiment 2, training proceeds as long as the RMSD between judgment and criterion does not fall below .8, as computed across the last 11 trials. The stop criterion is an attempt to assure that both the observation and intervention groups attained the same accuracy in training.

The hypothesis addressed is therefore that, at least, one factor that explains the improved learning with active intervention is that it provides more beneficial circumstances for abstracting knowledge in the form of abstract cue-criterion relations rather than exemplars, because intervention allows for more controlled experimentation. We expected that (a) interveners should need fewer trials of training to reach the stop criterion and (b) that while the observation group should be relatively more dominated by exemplar memory, the intervention group should reveal a shift toward cue abstraction. In addition, we predicted that performance would improve when training and test conditions were the same.

Method

Participants. Forty-eight undergraduate students (38 women, 10 men) from Uppsala University volunteered. All received payment of approximately 80 SKr (US$10) or course credit. The mean age of the participants was 25.5 years (SD = 5.66; range: 19–44). All participants were tested individually.

Materials and procedure. For the observers, each learning trial consisted of the presentation of text descriptions of a fictitious death bug species with four binary attributes (long or short legs, green or brown back, long or short nose, spotted or unspotted fore back). Each cue was presented by its text label (e.g., leg length) and with two labels representing the binary values of the cue (e.g., long, short). The actual cue value of a certain example was highlighted, whereas the other cue value was shaded. Five exemplars were omitted from training. The omitted bugs were Exemplars 1, 5, 6, 7, and 16 (see Table A2 of Appendix A). All participants were shown the exemplars in a new and independent random order.

The participants answered the question *What is the toxicity of this bug?* and made a numerical response (% toxicity) on each trial. Feedback about the correct toxicity followed the response and remained on the screen until the participant clicked to advance to the next trial. A minimum of 45 trials was completed and then training continued until the participant had satisfied a predetermined learning criterion (an average RMSD of 0.8 or less between judged and actual toxicity over the preceding 11 trials).

Interveners saw the same screen layout as observers, but rather than seeing predetermined configurations of attributes interveners selected four attributes on each trial (e.g., brown back, long nose, short legs, spotted fore back) by clicking on the cue value they desired. On each trial they were asked, for example, to *Create a bug that has toxicity 57%*. After selecting all four features and clicking *create*, feedback on the toxicity of the created bug was given. The program did not allow participants to create the five omitted exemplars, ensuring that in both conditions only the old exemplars in Table 1 could be seen.

Consistent with the observation condition, participants completed a minimum of 45 trials and then continued until achieving the same predetermined criterion. Following learning, half of participants from both groups received an observation test phase in which each of the 16 exemplars was observed and judged twice in a random order. The other half from both groups received an intervention test phase in which the task was to create a bug of a given toxicity. In both test conditions, the participants received no feedback for the 32 test trials.

Results

Performance. The results of Experiment 2 are summarized in Table 2. In Experiment 2 each participant received a different number of training trials depending on how fast they reached the stopping criterion of a maximal RMSE of .8 across the last 11 trials. Therefore, we only computed the experimentation index across the first 45 training trials that were performed by all participants. The mean experimentation index was 2.12 in the intervention condition and 1.86 in the observation condition, a difference that is statistically significant, $F(1, 46) = 20.79$, $MSE = .04$, $p = .000$. A two-way ANOVA on RMSE with training condition (observation vs. intervention) and test condition (observation vs. intervention) as between-subjects factors produces no statistically significant effect of test, $F(1, 44) = .39$, $MSE = .45$, $p = .536$, no significant interaction, $F(1, 44) = .98$, $MSE = .45$, $p = .328$, but a marginally significant effect of training, $F(1, 46) = 3.32$, $MSE = .45$, $p = .08$. Observation in training produced more accurate judgments than intervention (a RMSE of 1.21 vs. 1.56). The number of training trials needed to reach the training criterion was similar, $F(1, 46) = .14$, $MSE = 2038.31$, $p = .72$.

The mean judgments in the training conditions, illustrated in Figure 4A and 4B (collapsed across test conditions), suggest more accurate judgment in the observation training condition. In both data sets there is what first appears to be an inability to linearly

Table 2
Means for All Dependent Measures in Experiment 2

Measure	Observation	Intervention
RMSE	1.21	1.56
Training trials	90.10	78.20
Representation index (RI)	−1.30	−1.53
EBM (RMSD)	1.50	1.79
EBM (r^2)	0.73	0.48
CAM (RMSD)	0.27	1.74
CAM (r^2)	0.85	0.53

Note. RMSE = root mean square error; EBM = exemplar-based model; RMSD = root mean square deviation between model predictions and data; r^2 = coefficient of determination, variance accounted for by the model; CAM = cue abstraction model.

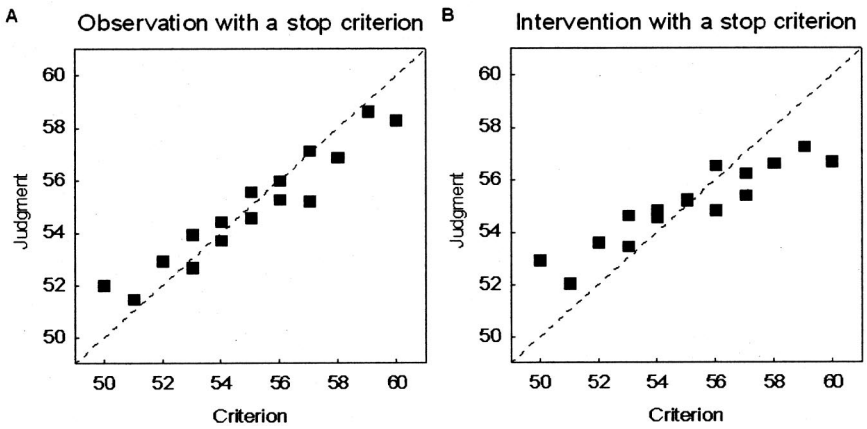

Figure 4. Data from Experiment 2. A: Mean judgments in the observation training condition collapsed across test conditions plotted against the criterion. B: Mean judgments in the intervention training condition collapsed across test conditions plotted against the criterion.

extrapolate for extreme exemplars, something that would support exemplar memory (see Figure 1). Closer scrutiny, however, reveals that these effects are explained mainly by reliance on cue abstraction, where the cue weight for the least important cue (C_4) has been erroneously estimated to a negative weight (see the *Model fit* subsection below). This is evident already from Figures 4A and 4B; there are large differences between the two old exemplars with Criterion 53, even though both of these exemplars were present in the training phase. This effect is explained by an erroneously estimated negative weight for Cue 4.

The correlation between the experimentation index and RMSE was $r_{48} = .19$, $p = .20$. The difference between this correlation and the correlation in Experiment 1 ($r_{32} = -.48$) is significant at $p = .0036$. The negative correlation between the intervener's experimentation index and RMSE in Experiment 1—suggesting that more active experimentation leads to more accurate judgment—has disappeared in Experiment 2, and the correlation is now (nonsignificantly) positive. The lack of a significant Training × Test Conditions interaction in regard to RMSE (accuracy), suggests that there were no differences between the groups with same training and test and the groups with different training and test.

Representation. A two-way ANOVA with RI as the dependent variable and training condition and test condition as between-subjects factors produced no statistically significant effects (all $p > .20$). In both the intervention and the observation training conditions, RI is statistically distinct from zero, but to large extent this is caused by many participants having learned the weight for Cue 4 in the erroneous negative direction.

Model fit. For each participant, the best-fitting parameters for each model were ascertained from the training phase and these parameters were then used to predict how the participant should perform in the test phase with all 16 exemplars (see Appendix A). A split-plot ANOVA with training and test condition (observation vs. intervention) as between-subjects factors and model (cue abstraction vs. exemplar memory) as within-subjects factor yielded two significant main effects (training and model) and one significant interaction between training condition and model, but it is the interaction, $F(1, 44) = 11.74$, $MSE = .098$, $p = .001$, that is the key effect. With observation training cue abstraction provides superior fit, but in the intervention training condition both models show the same rather poor fit.

Discussion

In contrast to Experiment 1, Experiment 2 revealed no benefit in learning for participants who learned the task by active intervention, despite the fact that intervention demonstrably increased the rate of experimentation in both experiments. Instead, there was a strong tendency toward more accurate judgments with passive observation. Moreover, the significant negative correlation between RMSE and the experimentation index in Experiment 1—suggesting that more experimentation improved judgment accuracy—disappeared and even turned into a nonsignificantly positive correlation in Experiment 2. It is important to note that this pattern of results was found even though the only difference between the tasks used in the two experiments was the use of binary rather than continuous cues.

Model fits failed to provide any evidence for a shift from exemplar memory to cue abstraction. Indeed, there was very little evidence for the exemplar memory model in either the observation or intervention group. The pattern of results are thus rather puzzling: With regard to the model fit, to the extent that either model dominates, it appears that cue abstraction provides a better overall fit, but even with this superiority for cue abstraction, no advantage is conferred by intervention. This pattern of results—active intervention with the stimuli producing poorer judgments than those made during passive observation—contrasts with the results of previous studies (Gopnik et al., 2004; Klayman, 1988; Lagnado & Sloman, 2004; Steyvers et al., 2003) and those of Experiment 1. This could indicate a limiting condition for the improvement in learning from intervention observed in previous studies. Perhaps the beneficial effects of intervention in multiple-cue judgment are limited to continuous cue environments.

An alternative possibility—addressed in Experiment 3—is that the training implied by the stop criterion used in Experiment 2 is too short for the performance difference and the representational shift to materialize. It is possible that in the intervention condition

the participants begin by using more or less fragmentary exemplars, and as training proceeds cue abstraction is slowly improved by more extensive experience with intervention. Several researchers have proposed that the early stages in learning involve a shift from rule-based processing to more instance-based processing (Johansen & Palmeri, 2002; Logan, 1988; Nosofsky & Palmeri, 1997). In the current context, the idea would be that in the early stages of training interveners rely on simple rules inferred from fragmentary knowledge of exemplars and as more complete exemplar knowledge is stored (through extended training) more complex rules about stimulus structure can be inferred through intervention, finally resulting in performance best fit by a cue-abstraction model. In contrast, observers who are denied the opportunity to "experiment" achieve complete exemplar knowledge (through extended training) but are unable to infer the more complex cue-criterion rules.

Experiment 3: Binary Cues and Extensive Training

We conducted Experiment 3 to further investigate whether active intervention with stimuli in training produces a representational shift between different representations. The training phase in earlier experiments with this paradigm have been 220 trials long (Juslin, Jones et al., 2003; Juslin, Olsson, & Olsson, 2003), and it is possible that this number of trials is necessary to see the shift from exemplar memory to cue abstraction.

The aim of Experiment 3 is therefore to investigate whether increasing the number of training trials would increase the difference between the two learning conditions, with regard to performance and the dominating representation. In addition, Experiment 3 serves the complementary purpose of collecting more data on one intriguing aspect of Experiment 2; the marginal deterioration in performance for interveners. Manipulation of observation versus intervention at test produced no substantial effects in Experiment 2, so we used observation in the test phase for both conditions in Experiment 3.

Method

Experiment 3 follows the same design and procedure as Experiment 2. The difference between the two experiments is the number of learning trials. Rather than continuing until reaching a predetermined learning criterion, all participants completed 220 trials in the training phase, ensuring that all 11 exemplars were presented (observation) or asked for (interveners) 20 times. Twenty-four undergraduate students (14 women, 10 men) from Uppsala University took part and were rewarded in the same manner as in Experiment 2. The mean age of the participants was 25.71 years (SD = 4.85; range: 20–45).

Results

The data in Experiment 3 were analyzed in the same way as those of Experiment 2. Benefiting from the fact that the experimental procedures, materials, and participant populations are the same in Experiments 1, 2, and 3, we make the comparison between the experiments to investigate the effects of task structure and increased training at the end of this *Results* section.

Performance. As in Experiment 1, the experimentation indices were computed for separate blocks of trials in the training phase. The indices for 11 blocks of 20 successive trials were entered as the dependent variable in an ANOVA. The ANOVA showed a significant main effect of condition, $F(10, 242) = 34.5$, $MSE = .08$, $p = .000$, a nonsignificant main effect of block, $F(10, 242) = .64$, $MSE = .08$, $p = .78$, and a nonsignificant interaction, $F(10, 242) = .67$, $MSE = .08$, $p = .75$. As illustrated in Figure 5, in the observation condition the experimentation index is flat at its random-base level throughout training; for the intervention condition it stays higher than this base level throughout training, with the highest value in the first block.

Figures 6A and 6B present mean judgments for each exemplar plotted against the criterion in both conditions of Experiment 3. As in Experiment 2, judgments by observers appear closer to the identity line and there is an apparent inability to extrapolate; although, as in Experiment 2, this seems to be primarily explained by cue abstraction with incorrect and negative cue weights for Cue 4 (e.g., the difference for the two old exemplars with Criterion 53 is larger than for the old–new comparisons at 55, 56, and 57; see Table A2 in Appendix A). A one-way ANOVA on the RMSE shows a marginally significant difference, again favoring passive observation over active intervention, $F(1, 22) = 3.22$, $MSE = .428$, $p = .09$. The correlation between the experimentation index and RMSE was significantly positive ($r_{24} = .50$, $p = .014$), suggesting that more experimentation was associated with poorer judgment accuracy. This correlation is significantly different from the negative correlation ($r = -.48$) that was observed in Experiment 1 ($p = .001$). Thus, as in Experiment 2, there was a strong tendency for interveners to engage in more active experimentation and fewer accurate judgments, thus their experimentation correlated with poorer judgments.

Representation. There was no significant difference in the RI between the intervention and observation conditions of Experiment 3, $F(1, 22) = 1.22$, $p = .28$. The RI are presented in Table 3. As in Experiment 2, these negative RIs seem primarily explained by cue abstraction based on erroneous cue weights (see section on Model fit). Also as in Experiment 2, in the intervention condition the correlation between the experimentation index and the RI was positive and this time significant ($r_{24} = .69$, $p = .014$).

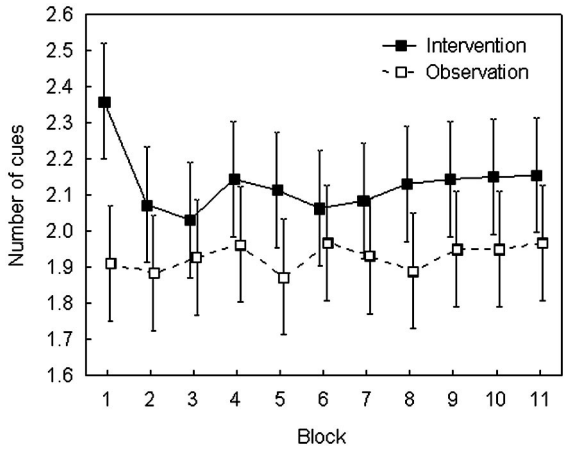

Figure 5. The mean experimentation index for the observation and intervention conditions of Experiment 3 plotted as a function of training block, where each block consists of 20 trails. The means for observation define the change of cues expected by chance.

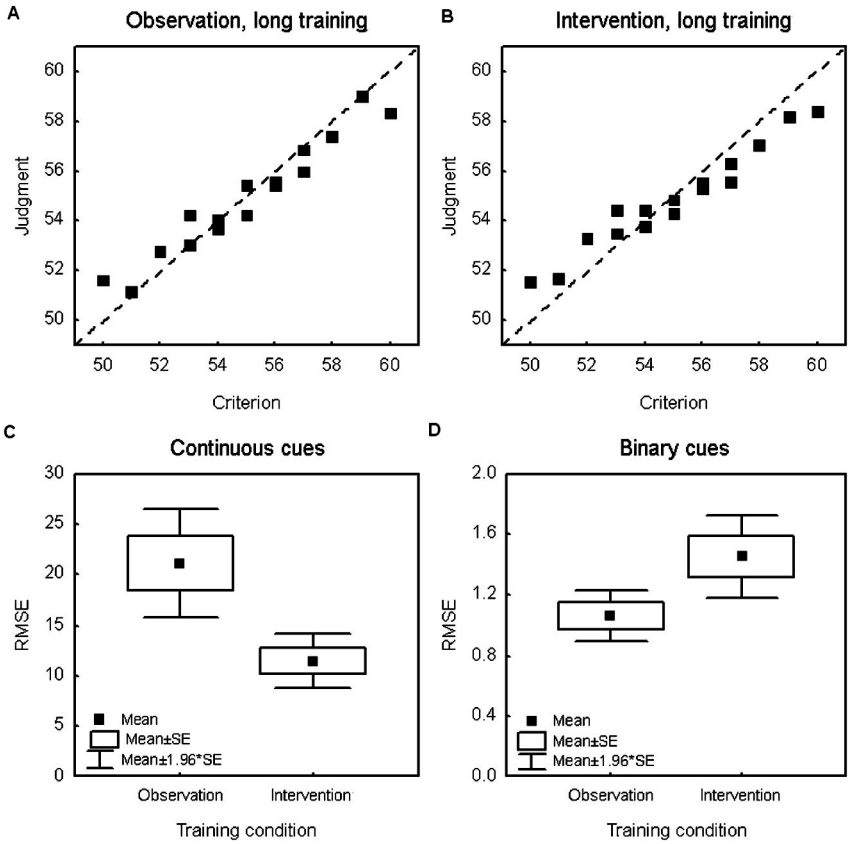

Figure 6. Panels A and B show data from Experiment 3 and Panels C and D show main empirical finding across Experiments 1, 2, and 3.A: Mean judgments in the observation condition plotted against the criterion. B: Mean judgments in the intervention condition plotted against the criterion. C: In a task with continuous cues, active intervention improves the accuracy of the judgments as compared with passive observation. D: In a task with binary cues, active intervention contributes to poorer accuracy of the judgments as compared with passive observation.

More experimentation and longer training increases the RI toward zero—suggestive of cue abstraction.

Model fit. Model fit was analyzed in the same way as in Experiment 2. A split-plot ANOVA with training condition (observation vs. intervention) as between-subjects factor and model (cue abstraction vs. exemplar memory) as within-subjects factor show a significant main effect of model, $F(1, 22) = 17.26$, $MSE = .14$, $p = .000$ and of training, $F(1, 22) = 10.16$, $MSE = .24$, $p = .004$, but no significant interaction, $F(1, 22) = .01$, $MSE = .14$, $p = .93$. Both models show better fit for the observation than the intervention condition, and the cue abstraction model shows superior fit in both of these training conditions. The correlations between the participants experimentation index and the RMSD of each model were $r_{24} = .58$ ($p = .003$) for the cue abstraction model and $r_{24} = .47$ ($p = .021$) for the exemplar model.

Table 3
Means for All Dependent Measures in Experiment 3

Measure	Observation	Intervention
RMSE	0.76	1.24
Representation Index (RI)	−1.46	−0.76
EBM (RMSD)	1.13	1.58
EBM (r^2)	0.80	0.55
CAM (RMSD)	0.67	1.14
CAM (r^2)	0.90	0.73

Note. RMSE = root mean square error; EBM = exemplar-based model; RMSD = root mean square deviation between model predictions and data; r^2 = coefficient of determination, variance accounted for by the model; CAM = cue abstraction model.

Analysis Over Experiments 1, 2, and 3

A two-way ANOVA across both Experiments 2 and 3 with RMSE as dependent variable and training condition (intervention vs. observation) and training criterion (stopping rule vs. extended training) as independent variables showed significant main effects of training condition, $F(1, 68) = 6.33$, $MSE = .44$, $p = .01$, and training criterion, $F(1, 68) = 5.6$, $MSE = .44$, $p = .02$, but no statistically significant interaction, $F(1, 68) = .15$, $MSE = .44$, $p = .7$. The more extensive training in Experiment 3 produced significantly more accurate judgments and—with this task involving binary cues and a continuous criterion—intervention contributed to poorer judgments as compared with passive observation of

the exemplars. A corresponding two-way ANOVA with RI as dependent variable and training condition and training criterion as independent variables shows no significant main effect or interactions (all $p > .2$).

The main empirical finding across the first three experiments is summarized in Figures 6C and 6D. In a multiple-cue judgment task with continuous cues (Experiment 1) active intervention facilitated learning and promoted more accurate judgments (Figure 6C), and the analysis of individual differences implied that experimentation was associated with more accurate judgments. In a multiple-cue judgment task with binary cues (Experiments 2 and 3), active intervention produced less accurate judgments than did passive observation (Figure 6D) and more experimentation was associated with poorer judgments.

Discussion

Overall, Experiment 3 provides further evidence that active intervention can actually instill poorer learning than passive observation in a task with binary cues, where more active experimentation contributes to poorer judgments. The increased training did not serve to separate the two conditions from each other and there were no signs of a shift from exemplar memory to cue abstraction. The overall superiority of cue abstraction was even clearer in this experiment as compared with the previous experiments, but this superiority does not appear to be strengthened by allowing the participants to actively intervene with stimuli.

Experiment 4: Binary Cues With All Exemplars Available

A factor that may have contributed to the interveners' poor performance in the binary task is restricting the types of exemplars that could be constructed in training. Recall that to allow the test of extrapolation and interpolation it was necessary to "hold back" five exemplars from the training phase. If a participant attempted to create one of these "restricted exemplars," an error message was presented. In the binary cue environment (Experiments 2 and 3), there are only 16 possible exemplars. When the five restricted ones are omitted, this leaves only 11 permutations to create. Given this low number, it is likely that participants encountered the error message with a high frequency. It seems plausible that being told repeatedly that one cannot make a particular exemplar would breed some frustration, which may well act to disrupt learning and impair judgment accuracy. This effect of the restricted exemplars may also explain the rather counterintuitive positive correlation between the experimentation index and RMSE in Experiments 2 and 3. It is important to note that in Experiment 1, in which there were 11^4 possible combinations, the likelihood of encountering a restricted exemplar was much lower and thus was much less likely to disrupt learning.

Experiment 4 investigates whether the poor performance for interveners in Experiments 2 and 3 is an artifact of holding back five exemplars in training. If receiving multiple error messages in training has a negative effect, for example, by breeding frustration that leads to less attention and motivation, this phenomenon would disappear when all exemplars are available in training. Because no exemplars were withheld in training, the models provide virtually identical predictions and no representation index or model fit was analyzed.

Method

Experiment 4 follows the same design and procedure as Experiment 3, with the difference that in Experiment 4 all 16 exemplars were used in both training and test. Thirty-two undergraduate students (24 women, 8 men) from Uppsala University participated. The mean age of the participants was 23.34 years ($SD = 2.71$; range: 19–32). Participants received either a movie ticket (approximately 80 SKr [US$10]) or course credits.

Results

An experimentation index was computed for separate blocks of trials in the training phase. The indices for 12 blocks of 20 successive trials were entered as the dependent variable in an ANOVA. The ANOVA showed a significant main effect of condition, $F(1, 360) = 55.15$, $MSE = .06$, $p = .000$, a nonsignificant main effect of block, $F(11, 360) = .32$, $MSE = .06$, $p = .98$, and a nonsignificant interaction, $F(11, 360) = .63$, $MSE = .06$, $p = .81$ (see Figure 7). A one-way ANOVA on RMSE showed a significant difference between the conditions, $F(1, 30) = 9.35$, $MSE = .89$, $p = .005$, in favor of observation. Figure 8 shows the judgment data from the test phase for both conditions. A correlation between experimentation index and performance (RMSE) showed a significant positive correlation ($r_{32} = .46$, $p = .008$), indicating, as in Experiment 3, that more experimentation is associated with poorer performance.

Disussion

In Experiment 4, we investigated if the poor performance for intervention in Experiment 2 and 3 can be explained by the fact that five exemplars were omitted in training. Therefore no exemplars were withheld in training in Experiment 4. The results show the same pattern as in Experiment 2 and 3, with even more distinct results. Observation has significantly better performance and there is a significant positive correlation between experimentation index and RMSE suggesting that more active experimentation leads to poorer performance.

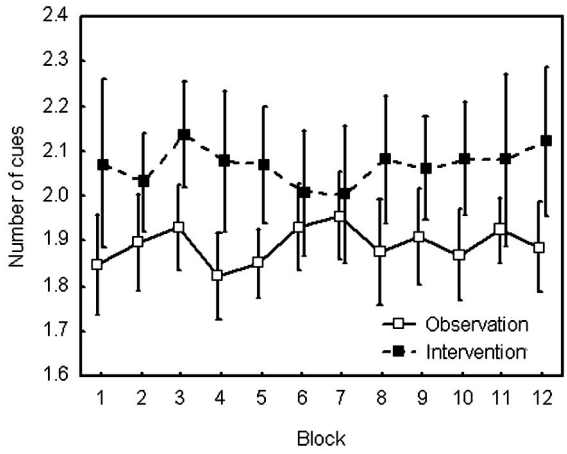

Figure 7. The mean experimentation index for the observation and intervention conditions of Experiment 4 plotted as a function of training block, where each block consists of 20 trails. The means for observation define the change of cues expected by chance.

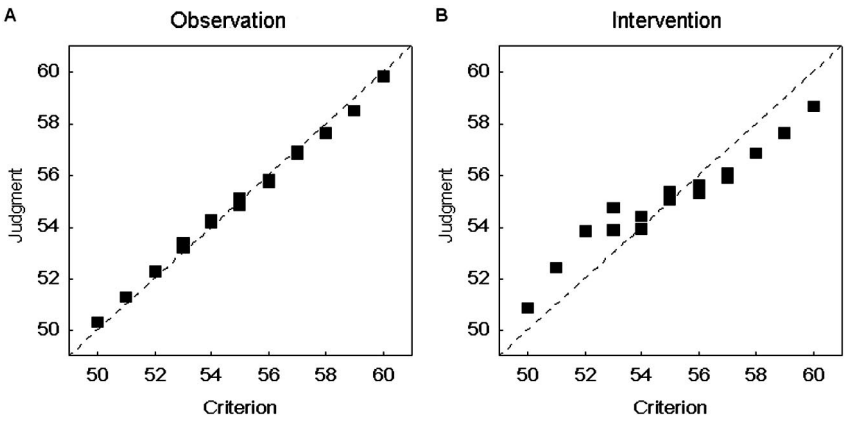

Figure 8. Data from Experiment 4. A: Mean judgments in the observation condition plotted against the criterion. B: Mean judgments in the intervention condition plotted against the criterion.

General Discussion

The standard paradigm in categorization and judgment research has been an "observation regime" in which participants passively observe stimuli (e.g., Medin & Schaffer, 1978; Nosofsky & Johansen, 2000). Recently, however, in both categorization research (e.g., Markman & Ross (2003); Ross, 2000) and causal reasoning research (e.g., Gopnik et al., 2004; Lagnado & Sloman, 2004) there has been increasing interest in the role that active strategies such as causal intervention play in learning. The standard finding accords with the intuition that learning is promoted by the opportunity to intervene with the system under study.

In four experiments, we tested a hypothesis about why intervention might confer a benefit for learning in a multiple-cue judgment task, and in doing so we discovered limiting conditions for the benefits of intervention. On the basis of Juslin, Olsson, & Olsson (2003) and Juslin, Jones, et al. (2003), we proposed that intervention might result in better learning than observation because intervention facilitates cue abstraction. In the environments we used, in which the to-be-judged criterion is well approximated by the linear, additive combination of cues, a cue-abstraction mechanism is a highly efficient strategy for learning (Juslin et al., 2004). We reasoned that providing the opportunity to conduct controlled "experimentation" in such an environment by, for example, systematically creating successive exemplars that differ with respect to a single cue, would offer an intervener a considerable advantage over an observer by enabling interveners to infer the orthogonal contribution of each cue.

Advantages and Disadvantages of Intervention

In Experiment 1, we found clear support for the hypothesized improvement. When both cues and criterion were pseudocontinuous (i.e., varied in 11 steps), intervention facilitated learning and resulted in more accurate judgments. We calculated an experimentation index—a measure of how many cue values were left unchanged between successive trials (Klayman, 1988). A value of 3 on this measure indicates holding all cues constant except one (i.e., *ideal experimentation*). We found that on average, even in the first block of trials, the index was only at 1.4, indicating less than perfect experimentation. Nevertheless, the value was significantly different from baseline (i.e., the random changes seen by observers) and decreased as the experiment progressed, suggesting improvements in learning (see Footnote 2). It is important to note that the correlation between experimentation index and RMSE was negative and significant, indicating that the greater the number of cues kept constant between successive trials, the lower the mean square deviation between the judgment and the criterion. Even this limited experimentation was sufficient for improving learning relative to observation.

By contrast, Experiments 2, 3, and 4 refuted the hypothesized improvement with intervention. In all three experiments, interveners produced fewer accurate judgments despite engaging in more active experimentation than observers. Experiment 4 investigated whether intervention could increase performance if the whole exemplar range were available in training, but in Experiment 4 the deviation between observation and intervention was even greater than in Experiments 2 and 3 in favor of observation. The significant negative correlation between experimentation index and RMSE found in Experiment 1 turned into a positive significant correlation in Experiments 3 and 4 (in Experiment 2 it was positive but not significant). In other words, more experimentation led to poorer judgments.

Evidence for Representational Shifts

In contrast to our hypothesis, the current experiments provided no evidence that the benefit of intervention is mediated by a representational shift from reliance on exemplar memory to cue abstraction. In Experiment 1, there was clear support for the cue-abstraction model in the intervention condition—in line with our hypothesis—but no clear support for either exemplar memory or cue abstraction in the observation condition. This latter finding is contrary to our hypothesis that exemplar memory should be more evident under observation training.

In Experiments 2 and 3, we examined whether the failure to find evidence for exemplar memory was a result of the continuous cue environment used in Experiment 1, acting as a strong trigger to analytical thinking. Experiments 2 and 3 used a binary cue environment but still failed to find clear support for exemplar memory. In both experiments, model fits were better for the observation

than the intervention conditions (indicative of less individual difference in the observation conditions), with cue abstraction showing a superior fit for both conditions. It is interesting to note that the data seemed to be best explained by a cue-abstraction model based on an erroneous estimation of the lowest weighted cue (Cue 4). Juslin, Jones, Olsson, and Winman (2003) demonstrated that a change from binary outcome feedback (e.g., toxic/not toxic) to continuous feedback (e.g., *toxicity was 57%*) led to a shift from exemplar memory to cue abstraction in the binary version of the bug task. In Experiments 2 and 3, although the feature information was binary, we provided continuous feedback about the toxicity of the bugs on each trial, something that may have reduced the potential to find exemplar effects.

A key difference was that in Experiments 2, 3, and 4 binary cues were used; in Experiment 1, cues that varied across 11 values (0–10) were used. Does this change from a continuous to a binary cue environment lead to a reversal in the fortunes of the interveners? Here we can only offer a tentative, but in our minds plausible, interpretation. A common conclusion in the recent literature is that the same judgment task is amenable to multiple processes that compete to control the response (Ashby et al., 1998; Erickson & Kruschke, 1998; Hammond, 1996; Sloman, 1996; E. E. Smith et al., 1998). According to this view, both cue abstraction and exemplar memory contributes to performance in all four experiments reported in this article. A second assumption is that although intervention affords a benefit for cue abstraction, its effect on the encoding of entire exemplars is detrimental. In Experiment 1, the number of exemplars is large and therefore exemplar memorization should be of less use, in favor of reliance on cue abstraction. The benefit for cue abstraction more than outweighs the deterioration in exemplar storage because exemplar storage only plays a marginal role.

In Experiments 2, 3, and 4, however, there are few exemplars and exemplar memorization is potentially a nontrivial contributor to performance, although on many trials the participants also used cue abstraction. Here the deterioration in the encoding of complete exemplars in the intervention condition may outweigh the benefit for cue abstraction, particularly if most participants are able to benefit from cue abstraction already in the observation condition, as suggested by the model fits. The behavioral dissociation across the continuous and binary tasks—that the same manipulation causes diametrically opposed effects on accuracy—indeed suggests that, to some extent at least, different processes operate in the two tasks. Obviously, however, the specifics of this interpretation of the poorer performance with intervention need to be further empirically validated.

Previous Comparisons of Observation and Intervention

Klayman (1988) found a clear and substantial advantage for interveners or experimenters over observers. The participants in Klayman's study had more scope for experimentation because they were in control of more aspects of each trial and were free to change all these aspects on any given trial, thus providing a rich environment in which to test hypotheses about cue–outcome relations. In our studies, interveners were provided with the criterion *toxicity* on each trial and asked to generate a bug with that toxicity. Providing the criterion in this way was necessary for maintaining consistency between the observation and intervention conditions, but it may have reduced the advantage conferred by intervention. Experiment 1, the experiment most similar to Klayman's experiments, however, replicates the main finding from Klayman: Interveners learn more efficiently. This suggests that it was not the more restrictive experimentation per se that produced the poor performance of interveners in Experiments 2, 3, and 4—rather it seems tied specifically to the binary cue environment.

The results from Experiments 2, 3, and 4, however, appear to contradict the results from a study by Lagnado and Newell (2003). The participants in that study learned to use a simplified drawing program where the positions of four binary buttons probabilistically determined which of two shapes were drawn (cube or cylinder). The interveners were not instructed to make a particular shape but were simply attempting to discover how each button contributed to the outcome. Observers viewed a randomly determined pattern of the four buttons before making a prediction. In a subsequent test phase, interveners made more accurate judgments of the likelihood that a given shape would be drawn than did observers, suggesting that intervention did confer an advantage in this binary environment.

The advantages for intervention in such causal reasoning tasks may be due to fundamental differences between the multiple-cue judgment and causal paradigms. One key difference is that in the causal literature that has examined intervention, the focus has been on learning causal relations that connect events (e.g., Lagnado & Sloman, 2004; Steyvers et al., 2003), but in multiple-cue judgment the task is generally to predict the criterion. Therefore, the causal reasoning task may more strongly invite representation in terms of causal models, that explicitly represent causal priority, rather than functional models that only state the covariation between cues and criterion. Moreover, although exemplar memory per se provides poor guidance in regard to causal reasoning, when the task is prediction it affords a viable alternative process. Future studies therefore need to identify more carefully what tasks and instructions promote strategies of learning that are benefited by the opportunity to intervene.

Conclusions

The main conclusions from this investigation are that (a) the opportunity to intervene confers an advantage in multiple-cue judgment when cues and criterion are presented on continuous scales but has a deleterious effect when binary cues are used; (b) participants do not spontaneously engage in controlled experimentation, rather they often change multiple-cue values between trials but that even this haphazard experimentation can lead to advantages in continuous cue environments; and (c) there is no evidence to suggest a shift from exemplar memory to cue abstraction, even when cue abstraction is facilitated by intervention.

References

Anderson, J. R., & Fincham, J. M. (1994). Acquisition of procedural skills from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1322–1340.

Anderson, N. H. (1981). *Foundations of information integration theory.* New York: Academic Press.

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.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications.* San Diego: Academic Press.

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.

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

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111,* 1–30.

Hammond, K. R. (1996). *Human judgment and social policy.* New York: Oxford University Press.

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

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief adjustment model. *Cognitive Psychology, 24,* 1–55.

Johansen, M. K., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology, 45,* 482–553.

Johansson, R., & Brehmer, B. (1979). Inferences from incomplete information: A note. *Organizational Behavior and Human Performance, 22,* 141–145.

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. (2004). *Additive integration of information in multiple cue judgment: A division of labor hypothesis.* Manuscript submitted for publication.

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

Klayman, J. (1988). Cue discovery in probabilistic environments: Uncertainty and experimentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 317–330.

Kuylenstierna, J. (1998). *Task information and memory aids in the learning of probabilistic inference tasks.* Uppsala, Sweden: Uppsala University.

Lagnado, D. A., & Newell, B. R. (2003). *Intervening versus observing in multiple-cue probability learning.* Paper presented at the Society for Judgment and Decision Making, Vancouver, British Columbia, Canada.

Lagnado, D. A., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 856–876.

Logan, D. G. (1988). Toward and instance theory of automization. *Psychological Review, 95,* 492–527.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin, 129,* 592–613.

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

Mill, J. S. (2002). *A system of logic: Rationcinative and inductive.*: Miami, FL: International Law and Taxation. (Original work published 1843)

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16,* 519–533.

Nilsson, Olsson, & Juslin. (2005). The cognitive substrate of subjective probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 600–620.

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. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–61.

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

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104,* 266–300.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101,* 53–79.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* New York: Cambridge University Press.

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1141–1159.

Ross, B. H. (1996). Category representations and the effects of interacting with instances. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1249–1265.

Ross, B. H. (2000). The effects of category use on learned categories. *Memory & Cognition, 28,* 51–63.

Ross, B. H., & Warren, J. L. (2002). Learning abstract relations from using categories. *Memory & Cognition, 30,* 657–665.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119,* 3–22.

Smith, E. E., Patalano, A. L., & Jonides, J. (1998). Alternative strategies of categorization. *Cognition, 65,* 167–196.

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.

Sobel, D. M. (2004). *Watch it, do it, or watch it done: The relation between observation, intervention, and observation of intervention in causal structure learning.* Manuscript submitted for publication.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27,* 453–489.

(*Appendixes follow*)

Appendix A

Task Structure in Experiments 1 and 2

Table A1
All Exemplars With Continuous Cues Values Used in Experiment 1

	Cue				Criterion			Cue				Criterion	
Exemplar	1	2	3	4	Value	Role	Exemplar	1	2	3	4	Value	Role
1	10	0	10	0	600	*E*	30	10	10	0	2	548	*T*
2	10	0	8	0	596	*E*	31	7	8	4	5	547	*T*
3	9	0	10	2	594	*E*	32	2	0	3	10	544	*O*
4	9	1	9	0	591	*E*	33	2	1	4	9	544	*N*
5	10	0	10	10	590	*T*	34	0	0	5	9	541	*T*
6	9	1	9	2	589	*T*	35	5	10	5	1	539	*T*
7	10	1	10	10	587	*T*	36	1	2	5	10	538	*T*
8	9	0	9	8	586	*O*	37	5	5	0	10	535	*T*
9	8	0	9	4	586	*N*	38	0	6	10	9	533	*O*
10	8	1	10	4	585	*T*	39	1	7	10	10	533	*N*
11	8	2	8	0	582	*T*	40	1	9	10	7	530	*T*
12	8	2	9	5	579	*T*	41	1	9	10	9	528	*T*
13	7	3	8	0	575	*T*	42	3	7	2	8	527	*T*
14	10	0	0	6	574	*O*	43	5	10	2	10	524	*O*
15	10	0	2	10	574	*N*	44	2	8	4	8	524	*N*
16	10	0	0	10	570	*T*	45	1	9	3	2	521	*T*
17	8	2	5	9	567	*O*	46	0	10	10	10	520	*T*
18	9	2	3	9	567	*N*	47	1	8	3	8	518	*O*
19	10	5	5	9	566	*T*	48	0	8	4	6	518	*N*
20	1	0	10	0	564	*T*	49	2	10	2	6	516	*T*
21	7	1	2	8	561	*O*	50	0	10	5	5	515	*T*
22	7	2	3	7	561	*N*	51	1	10	5	10	514	*O*
23	6	5	5	1	558	*T*	52	0	6	1	10	514	*N*
24	6	4	5	5	557	*T*	53	2	10	0	7	511	*T*
25	3	1	5	4	555	*O*	54	1	9	1	9	510	*T*
26	10	10	5	5	555	*N*	55	1	9	0	9	509	*E*
27	5	5	6	4	553	*T*	56	1	10	1	10	506	*E*
28	5	5	5	5	550	*T*	57	0	9	0	9	504	*E*
29	5	5	5	6	549	*T*	58	0	10	0	10	500	*E*

Note. Two cues are positively linearly related (Cues 1 and 3) and two cues are negatively linearly related (Cues 2 and 4). T and O refer to exemplars viewed under both training and test, where O signifies an old exemplar matched to a new exemplar. N and E are new exemplars presented only in the test phase.

Table A2
All 16 Exemplars Used in Experiment 2 and Their Binary Cue Values

	Cue				Criterion	
Exemplar	1	2	3	4	Value	Role
1	1	1	1	1	60	*E*
2	1	1	1	0	59	*T*
3	1	1	0	1	58	*T*
4	1	1	0	0	57	*O*
5	1	0	1	1	57	*N*
6	1	0	1	0	56	*N*
7	1	0	0	1	55	*N*
8	1	0	0	0	54	*T*
9	0	1	1	1	56	*O*
10	0	1	1	0	55	*O*
11	0	1	0	1	54	*T*
12	0	1	0	0	53	*T*
13	0	0	1	1	53	*T*
14	0	0	1		52	*T*
15	0	0	0	1	51	*T*
16	0	0	0	0	50	*E*

Note. T and O Refers to Exemplars Viewed both Under Training and Test, where O Signifies an Old Exemplar Matched to a New Exemplar. N and E are New Exemplars only Presented in the Test Phase.

Appendix B

Quantitative Implementation of the Models

In essence, when the participants make judgments of the continuous criterion the cue abstraction model suggests that they perform a mental analogue of linear multiple regression. For each cue, the weight ω_i (i = 1…4) is retrieved and the estimate of c is adjusted accordingly:

$$\hat{c}_{\mathrm{R}} = k + \sum_{\mathrm{i}=1}^{4} \omega_{\mathrm{i}} \cdot C_{\mathrm{i}}, \tag{B1}$$

where $k = 500 + .5 \cdot (100 - 10 \cdot \omega_i)$. If $\omega_1 = 4$, $\omega_2 = 3$, $\omega_3 = 2$, and $\omega_4 = 1$, Equations 1 and B1 are identical and the cue abstraction model affords perfect judgments in this task.

The exemplar-based model implies that the participants make judgments by retrieving similar exemplars from memory (Medin & Schaffer, 1978). When the exemplar model is applied to judgments of a continuous criterion, the estimate $\hat{C}_E$ of the criterion c is a weighted average of the criteria c_j stored for the J exemplars, where the probe-exemplar similarities $S(p, x_{\mathrm{j}})$ are the weights:

$$\hat{c}_E = \frac{\sum_{j=1}^{J} S(p, x_j) \cdot c_j}{\sum_{j=1}^{J} S(p, x_j)}, \tag{B2}$$

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 probe p and exemplar x_j. The similarity between the probe p and exemplar x_j is computed according to the generalized context model (GCM: (Nosofsky, 1984, 1986), a generalization of the original context model. The similarity $S(p, x_j)$ between a probe p and an exemplars x_j is found by transforming the distance between them.
According to GCM, the distance between a probe p and an exemplar j is,

$$d_{pj} = h[\sum_{M=1}^{M} w_{\mathrm{m}} |x_{\mathrm{pm}} - x_{\mathrm{jm}}|], \tag{B3}$$

where x_{pm} are the value of the probe and x_{jm} are the values of an exemplar on the cue dimension m, the parameters w_{m} are the attention weight associated with cue dimension m, and h is a sensitivity parameter that reflects the overall property of discrimination in the psychological space. Attention weights can vary between 0 and 1 and are constrained to sum to 1. The similarity $S(p, x_j)$ between a probe p and an exemplar j is assumed to be a nonlinear decreasing function of the distance (d_{pj}) between them,

$$S(p, x_j) = e^{-d_{pj}}. \tag{B4}$$

Received May 25, 2005
Revision received September 20, 2005
Accepted September 29, 2005 ■