Journal of Experimental Psychology:
Learning, Memory, and Cognition
1995, Vol. 21, No. 4, 815–836

Effect of Memory Decay on Predictions From Changing Categories

Stuart W. Elliott and John R. Anderson
Carnegie Mellon University

In contrast to the static categories assumed in most categorization experiments, many real-world categories undergo gradual and systematic change in their definitions over time. Four experiments were carried out to study such category change. In these studies, participants successfully adjusted as category change occurred, but also showed a lingering and cumulative effect of past observations. The participants' performance was closely modeled by incorporating memory decay for past observations into J. R. Anderson's (1990, 1991) rational categorization algorithm and into a version of R. M. Nosofsky's (1986) exemplar categorization model. The resulting models suggest that the decay function is closer to a power law than to an exponential and that decay occurs both by item and by time, with the item decay being stronger than the time decay. The finding of power law decay gives additional support to claims that exemplar memories are used in categorization.

It is not possible to step twice into the same river.—Heraclitus

From the examples typically discussed in the categorization literature, it can seem that most category definitions are relatively fixed: Dogs don't switch from barking to meowing, and poodles don't mutate into German shepherds. The evolutionary change that does occur in such categories is glacial with respect to the pace of category formation and use that occurs during a human lifetime. However, the stability at this relatively abstract category level belies the extensive change that occurs at the category level specifying individuals: Poodles in general may not change much over time, but one individual poodle will grow and change, with a steady mutation in appearance and behavior over time.[1]

At more aggregate category levels there is also change that unfolds fast enough to be experienced by humans. It has become a truism to note the extensive technological and social change that occurs during a single person's lifetime. Cars, televisions, and computers have all evolved significantly over the past few decades, as have fashions in clothing, the structure of our cities, and the roles of women and minorities in our society. Like the categories specifying individuals, these more aggregate categories require adjustment in their definitions over time.

Categories that gradually change expose models of categorization to a somewhat different set of requirements than static categories do. In particular, changing categories highlight the interaction between new observations and existing category definitions. With fixed underlying categories, successive observations do not differ systematically from earlier observations, and in consequence it does not matter much whether the later observations are being used to update the category definitions. In contrast, with changing categories the later observations are systematically different from the earlier observations so that it makes a significant difference whether these later observations have a large or small impact on the category definitions, or whether they have none at all.

Models of categorization differ in the relative weight they put on new and old observations. Many symbolic categorization algorithms, including prototype, feature frequency, and exemplar models (Estes, 1986; Smith & Medin, 1981), weight the information from all past observations equally. In contrast, many connectionist algorithms (Gluck & Bower, 1988; Rumelhart, 1989; Rumelhart, Hinton, & Williams, 1986) incorporate the assumption that greater weight should be placed on more recent observations and that the relative weight of older observations should decline exponentially as successive observations are added. With the delta rule, decay is exponential because the information from all relevant past observations is decayed by the same fixed factor in relation to the information from a new observation. Exponential decay of past observations has also been incorporated in exemplar models (Estes, 1994; Nosofsky, Kruschke, & McKinley, 1992).

A categorization algorithm incorporating exponential decay will adjust more quickly to changing category definitions than will an algorithm that weights past observations equally, because the exponential decay will put relatively more weight on the more relevant observations. With equal weighting, adjustment to a change can occur only as the number of postchange observations becomes large relative to the number of prechange observations. Thus, algorithms using exponential decay have a built-in feature for adjusting to changing category definitions, whereas algorithms that weight observations equally

Stuart W. Elliott and John R. Anderson, Department of Psychology, Carnegie Mellon University.

This work was supported by Training Grant 5 T32 MH19102-03 from the National Institute of Mental Health and by Grant 91 08529 from the National Science Foundation. We would like to thank Jon Fincham and Ken Koedinger for helpful comments on this article. In addition, we would like to thank Jon Fincham for his assistance in programming the experiments.

Correspondence concerning this article should be addressed to Stuart W. Elliott, who is now at the Graduate School of Industrial Administration, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213. Electronic mail may be sent via Internet to stuart.elliott@cmu.edu.

[1] We are grateful to several reviewers for pointing out that Barsalou's (1989) work on ad hoc and context-dependent categories is a notable exception to the implicit assumption of stability in much of the categorization literature.

require a large number of observations before showing much adjustment if they already have much experience before the change occurs. This contrast raises the question of whether exponential decay of past observations is consistent with people's ability to adjust to changing categories.

Another possible form of adjustment, suggested by the literature on memory retention over time, is that the impact of past observations will decline according to a power law function. A theoretical analysis of the potential relevance of past observations to current goals suggests that memories should decay according to a power function of time (Anderson, 1990). This theoretical analysis agrees with environmental estimates of the actual decline in relevance of past observations over time (Anderson & Schooler, 1991) and with empirical work on human and animal memory retention over time (Wickelgren, 1974; Wixted & Ebbesen, 1991). There is no reason to think that this analysis of potential relevance should be any different when the information in memory is aggregated into categories than when it is retrieved separately as memories (Elliott, 1991). Thus, an adaptive approach to categorization suggests that the relative weight of past information should decline according to a power function.

Equal weighting and exponential decay are convenient to use in categorization models because they both allow the information from past observations to be recorded with summary statistics for each feature of each category. For equal weighting, a new mean is formed by taking a weighted average of the old mean and the new observation, with relative weights of $n/n + 1$ and $1/n + 1$, respectively, where n is the number of past observations. Similarly for exponential decay, a new mean is formed using relative weights of d and $1 - d$, where d is the decay factor. Equal weighting requires two summary statistics—the current mean and number of past observations—whereas exponential decay requires only one summary statistic for the current mean (assuming the decay factor is a fixed system parameter). In contrast, power law decay requires that all past observations be separately represented because the weights of all past observations are decaying at different rates. It is possible to approximate power law decay with exponential decay to a nonzero asymptote, which does not require separately representing all past observations.[2] However, given both the finding of power law decay in the memory retention literature and the natural representation of that decay using individual observations, a finding of power law decay would add weight to the arguments that there is an exemplar basis for categorization (e.g., Medin & Schaffer, 1978; Nosofsky, 1986).

Related Literatures

The categorization literature has investigated the effect of abrupt changes in category definitions. The Wisconsin Card Sorting Test involves category definitions that change from one feature to another as soon as participants have mastered the first definition. An inability to adjust is used as an indicator of impairment, because normal participants are able to accommodate themselves to successive shifts of category definitions (Berg, 1948; Grant & Berg, 1948; Robinson, Heaton, Lehman, & Stilson, 1980). The ability to adjust relatively quickly suggests that participants give more weight to more recent observations. Estes (1989) reported a similar experiment in which two category definitions were switched. In that experiment, adjustment was the same whether the switch occurred after 60 or after 180 training trials, suggesting again that participants place most of their attention on the more recent observations. These experiments with abrupt category change suggest a recency effect in categorization, but this inference is problematic because the abruptness of the change may cause the original categories to be abandoned rather than updated.

Several researchers have looked at the detailed course of learning for static categories with results that suggest a recency effect for changing categories. Busemeyer and Myung (1988) found that more recent observations were weighted more heavily when participants were asked to construct their estimates of category prototypes after successive observations. Nosofsky et al. (1992) fitted an exemplar model to learning data and found an exponential decay factor of about 0.98 per trial. Unlike connectionist models that require some exponential decay in order for learning to take place at all, Nosofsky et al.'s exemplar model allows either exponential decay or equal weighting, so that a finding of a decay factor less than 1.0 is an indicator of a recency effect. Although a decay factor of 0.98 may seem close to 1.0, it implies that more than half the total weight is being placed on the most recent 35 exemplars in an experiment with 240 observations.

Work on adjustment to change has been done in literatures related to but different from categorization. The cue probability learning literature concerns the ability to learn to predict a target dimension from several cue dimensions. This literature tends to look at tasks involving continuous linear relations between cue and target dimensions, whereas the categorization literature tends to look at tasks involving isolated clusters of correlated dimensions that correspond to separate categories. Within the cue probability learning literature, there has been work on the ability of participants to adjust to shifting cue validities (e.g., Ruffner & Muchinsky, 1978). Overall, this literature indicates that participants become less accurate after cue validity shifts but that they are relatively successful in redirecting their attention toward those cues that have become more predictive of the target dimension. In the studies we have seen, there was only a single cue validity shift.

A related adjustment task has been investigated in the belief updating literature (see Hogarth & Einhorn, 1992, for a review). This literature is concerned with understanding the incremental effects on participants' beliefs caused by the addition of successive pieces of information. The information can be presented to the participants in an ordered fashion, with largely negative items followed by largely positive items, or vice versa. Although it is somewhat unnatural, one way to describe this information ordering is that a change occurs, from negative to positive or the reverse; participants who show a recency effect in their beliefs have adjusted to this change in information, whereas participants who show a primacy effect have not. Hogarth and Einhorn (1992) summarized a large

[2] We thank John Wixted for pointing out this approximation. Exponential decay to a nonzero asymptote can be represented as a weighted combination of exponential decay to zero and equal weighting.

number of studies of the effects of presentation order. With short series of simple pieces of information, participants showed a primacy effect if they did not have to give a response until the end of the series, but they showed a recency effect if they had to give a response after each new information item. With more complex pieces of information, participants tended to show a recency effect even if they did not have to respond until the end of the series. With longer series of information, participants tended to show a primacy effect even if they had to respond after each new information item.

The problem of category change addressed here is different than the problem addressed by the conceptual change literature (Carey, 1985; Keil, 1989), although the labels are close enough to be confusing. The conceptual change literature concerns a series of conceptualizations that a learner may go through in learning about a static external pattern. The change that is the focus of that literature is internal to the learner. In contrast, the category change that we are concerned with is originally external to the learner, because the pattern that the learner is trying to learn is actually changing over time.

To the extent that existing work looks at changing categories, people seem to show relatively successful adjustment to change. However, the existing literature does not look extensively at people's ability to adjust to categories that show continual and gradual change. The contrast between gradual and abrupt change is critical, because abrupt change may lead to the creation of new categories rather than the updating of old categories. Furthermore, where the existing literature does show a recency effect in category-based judgments, no attempt has been made to determine the exact form of the strength decay for past observations.

We turn now to the first of four experiments, whose purpose is simply to look at the general course of adaptation to such gradually changing categories. After establishing that people's performance is biased toward their more recent observations, we describe two models incorporating memory decay that can produce a recency effect. In Experiment 2 we then present two methods for investigating the exact form of the decay of information from past observations. One method involves seeing how participants' behavior changes after an inserted delay, and the other method involves fitting the decay function of the two models to a complex set of changes. In Experiment 3 we investigate further the effect of an inserted delay, and in Experiment 4 we investigate further fitting the decay function to a complex set of changes.

Experiment 1

Experiment 1 was designed simply to see how successfully participants can follow a series of changes in two category definitions. The experimental task involved learning how to predict the missing dimensions of the stimuli. Participants made predictions for a series of stimuli, receiving correction after each. Unlike most categorization studies, the participants' ability to group the stimuli into the two underlying categories was not measured directly. However, because of the categorical structure of the stimuli, participants could not do better than chance without being sensitive to that categorical structure.

Method

Participants. The participants were 20 young adults from the Carnegie Mellon University community, who received either course credit or $5 for their participation.

Stimuli and design. The stimuli were pairs of ovals presented on a Macintosh IIci computer with a monochrome two-page monitor. The height and the width of the 2 ovals were treated as four dimensions. The height and width of 1 of the ovals were both fixed, whereas the height and width of the other oval both changed in a series of steps over the course of the experiment. The two fixed dimensions provided unambiguous information about category membership. The change on the other two dimensions took place over a series of five blocks of observations. Within each block, there were 16 stimuli from each of the two categories, for a total of 160 stimuli for the entire experiment. Figure 1 shows the mean values for the two categories on Blocks 1, 3, and 5.

As the figure shows, the change was defined so that the categories exchanged one of their ovals: Category A started out with a tall thin oval and ended up with a short wide oval, whereas Category B did the reverse. The location of the changing oval was counterbalanced, so that half of the participants saw the pairs as in Figure 1, with the changing oval on the bottom, and half saw the pairs with the changing oval on the top. The order of the stimulus series was also counterbalanced, with half of the participants seeing the series in the forward direction and half in the reverse direction.

The dimensions were coded on a unit interval, where the full value of 1 corresponded to a maximum height or width of 3.15 in. (8.0 mm) on the computer screen. For the fixed oval, Category A had a mean height of 0.6 unit and a mean width of 0.8 unit, and Category B had a mean height of 0.4 unit and a mean width of 0.2 unit. For the changing oval, the mean height ranged between 0.4 unit and 0.6 unit, with intermediate stops at 0.45, 0.5, and 0.55 unit, and the mean width ranged between 0.2 unit and 0.8 unit with intermediate stops at 0.35, 0.5, and 0.65 unit.

The actual observations were distributed around each of the block means with a variance of 0.005, resulting in a standard deviation of about 0.07. Within each block of observations, 16 stimuli were created for each category such that the values on each dimension had the

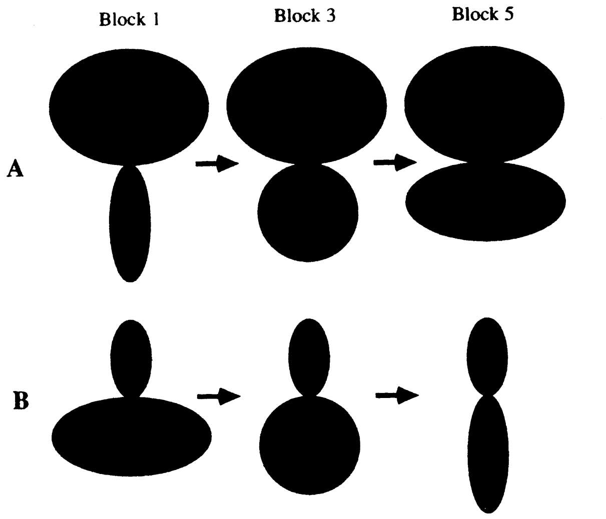

Figure 1. Mean values for stimuli used in Experiment 1, for Categories A and B on Learning Blocks 1, 3, and 5. Stimuli are shown in the forward direction, with the changing oval on the bottom.

appropriate distributions. The 16 stimuli for the two categories were then randomly mixed within each block, resulting in a single stimulus series with 160 items. The same single stimulus series was used for all participants. The variance was chosen to be large enough to mask the changes between each block but small enough that participants would feel relatively confident about predicting values on each dimension. It was important to mask the change enough that participants would respond on the basis of their actual observations; an unmasked change might be noticed by some participants who could then start to anticipate future changes, rather than rely on their past observations. For the changing width, each incremental step increased the mean by about two standard deviations, whereas for the changing height each step increased the mean by about 0.7 standard deviation.

For each stimulus, participants were given one of the ovals to use to predict the other oval. On half of the trials they used the fixed oval to predict the changing oval, and on the other half they used the changing oval to predict the fixed oval. Each stimulus was initially presented with the oval to be predicted replaced by an anchor oval in the middle of the distribution, with height and width of 0.5 unit. The four dimensions of the stimulus were manipulated by using a mouse to control four pairs of buttons on the screen, with one button to increase and one to decrease each dimension. For each stimulus, only two pairs of buttons were activated, corresponding to the two dimensions to be predicted.

When the participant had completed each prediction, the correct oval pair was shown along with a superimposed outline of the participant's prediction. The drawing was accompanied by an error-units score that ranged from 0 to 100. The score was obtained by taking the square root of the mean squared error of the two predicted dimensions and multiplying it by 100. Good scores were followed by comments: "Good" for scores from 8 to 14, "Very good!" for scores from 4 to 7, and "GREAT!" for scores from 0 to 3.

Procedure. Participants received both verbal and on-line instructions that their task was to "learn to make predictions about pairs of ovals" and that this would involve figuring out "what a correct oval pair should look like." None of the instructions mentioned that the oval pairs were structured around categories, and none of the instructions suggested that the criteria for a correct oval pair might change over the course of the experiment. We gave participants a brief practice session so that they would be able to use the controls to change the shapes of the ovals and to interpret the corrections they received after each prediction. Once they were set up on the computer, participants proceeded automatically from the on-line instructions to the practice session and then to the prediction task. We left participants alone for the entire computer session, but interrupted them briefly after the initial 5 min to ask if they understood the instructions and the controls. Participants had unlimited time to make the prediction and view the correction for each stimulus. After the computer session, we showed participants (simultaneously) 12 pairs of ovals and asked them to rate their typicality on a scale from 1 (*never appears*) to 5 (*very often appears*). The stimuli used in this rating were the mean oval pairs for both categories at Blocks 1, 3, and 5, along with the same 6 pairs shown upside down.

Results and Discussion

Figure 2 shows the course of learning for all participants over the five blocks of the experiment for the two changing dimensions. In this figure, the actual predictions have been recoded to show the percentage of change they indicate, according to the appropriate dimension and category. For example, the width of the changing oval for Category A for participants going in the forward direction increases from 0.2 to 0.8 unit. An observation of 0.35 unit on this dimension for Category A would be coded as 0.25, because it represents 25% of the distance from 0.2 to 0.8 unit. To give another example, the height of the changing oval for Category A for participants going in the forward direction decreases from 0.6 to 0.4 unit. An observation of 0.45 unit on this dimension for Category A would be coded as 0.75, because it represents 75% of the distance from 0.6 to 0.4 unit. After recoding, it is easy to compare the predictions with the changing mean of the observations, which has values 0, 0.25, 0.5, 0.75, and 1.0 unit over the course of the 5 blocks.

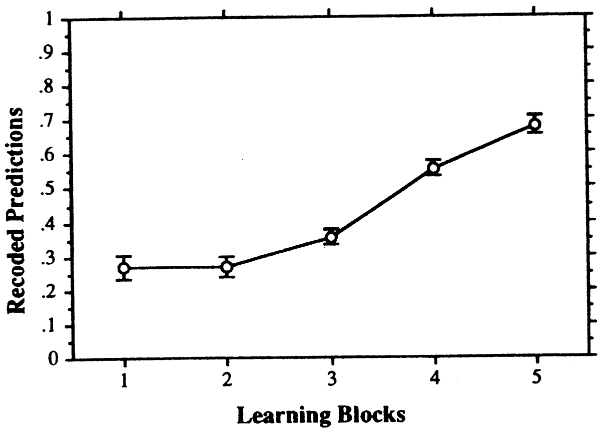

Figure 2. Predictions on changing dimensions for Experiment 1. Predictions were averaged over categories and dimensions, after being recoded so that the true mean of each category of each dimension moved from 0 on Block 1 to 1 on Block 5. Error bars represent standard errors.

Two conclusions can be drawn from Figure 2. First, participants adjusted more quickly than they would have if they had weighted all past observations equally. If participants had weighted all observations equally, then by the end of Block 5 their recoded predictions would have reached 0.5 unit, with the average over Block 5 being somewhat less than 0.5 unit. However, the mean of the participants' recoded predictions in Block 5 was 0.679 unit, which is significantly greater than 0.5 unit, $t(19) = 6.05, p < .001, MSE = 0.0174$. This higher mean indicates that participants were giving more weight to the more recent observations, which had higher values than did the earlier observations.

The second conclusion that can be drawn from Figure 2 is that participants adjusted less quickly to the changes than they would have if they had paid attention to only the most recent observations. If participants had paid attention to only the most recent observations, then their predictions would have closely followed the progress of the true values, and their average recoded predictions over all five blocks would have been 0.5 unit. However, the mean of the participants' recoded predictions over all five blocks was 0.424 unit, which is significantly less than 0.5 unit, $t(19) = 4.23, p < .001, MSE = 0.00637$. This lower mean indicates that the participants' predictions were being held down by the continuing effect of earlier observations. This effect was particularly striking during Blocks 4 and 5. Note that during Block 1 the mean of the participants' recoded predictions was substantially above the true observation mean of 0. This is not surprising, because when participants started making predictions in Block 1 they knew nothing about the stimuli structure, and so their early predictions had to have been noisily distributed around the 0.5-unit midpoint of the interval.

The typicality ratings reinforce the conclusion that participants placed greater weight on their more recent observations. In the typicality ratings, there was a main effect of block, $F(2, 38) = 6.210, p < .01, MSE = 2.12$. The typicality ratings were taken at the end of the experiment, with the more recent Block 5 figures receiving a mean of 3.63, compared with the earlier Block 1 and Block 3 figures, which had means of 2.89 and 2.96, respectively.

Both the predictions and the typicality ratings made by participants were biased toward the more recent observations. These results cannot be accommodated by categorization models that weight all past observations equally. They require instead that more weight be placed on the more recent observations. We therefore turn next to two models of category-based predictions that incorporate memory decay.

Models of Category-Based Predictions Incorporating Memory Decay

Two models of recency-biased predictions were constructed by incorporating memory decay into different categorization models. The first model is based on Anderson's (1990, 1991) rational categorization algorithm, and the second is an exemplar categorization model based on Nosofsky's (1986) generalized context model. Both models are briefly reviewed below. Two categorization models were used here to deflect attention from the idiosyncrasies of the models themselves and focus it instead on the addition of memory decay to the models and on the form that the memory decay takes.

The central problem that any categorization algorithm faces is to determine the category of a new observation given its observable feature structure. In Anderson's (1990, 1991) rational categorization algorithm, the probability of the observation falling into any category k, conditional on its feature structure F, is calculated with Bayes's rule using the prior probabilities of each of the categories $P(k)$ and the conditional probability of observing F from each of the categories $P(F|k)$:

$$P(k|F) = \frac{P(k)P(F|k)}{\sum_k P(k)P(F|k)}. \tag{1}$$

The prior probability $P(k)$ is calculated both for the k categories that the algorithm has already created from its previous n observations and for the possible new category, $k = 0$, that could be created for the new observation:

$$P(k) = \frac{cn_k}{(1-c) + cn} \tag{2}$$

$$P(0) = \frac{(1-c)}{(1-c) + cn}, \tag{3}$$

where n_k is the number of observations in each category, and c is the coupling parameter, which is the probability that any two objects come from the same category. The derivations of these formulas can be found in Anderson (1990). The formulas require an a priori decision about how stringent to be in grouping items, ranging from putting all items into separate groups ($c = 0$) to putting all items into the same group ($c = 1$).

The conditional probability of observing the feature structure from each of the categories, $P(F|k)$, is calculated as the product of the feature probability densities of the individual i features, $f_i(F_i|k)$, on the basis of the assumption that all features are independent:

$$P(F|k) = \prod_i f_i(F_i|k), \tag{4}$$

where F_i refers to the ith feature of the feature structure F. Feature probability densities are used instead of feature probabilities because all four dimensions are continuous rather than discrete.

The feature probability densities for each dimension i, $f_i(x|k)$, are calculated according to a Bayesian analysis for estimating a normal distribution with unknown mean and variance. This analysis results in a t distribution for the probability density of the current observed value x that is a function of the prior mean and variance, μ_0 and σ_0^2, the confidence in the values for the prior mean and variance, λ_0 and α_0, and the mean and variance of the previous n_{ik} observations of dimension i for that category, $\bar{x}_{ik}$ and s_{ik}^2 (Anderson, 1991). Thus, the model begins with prior values for the unknown mean and variance of the normal distribution, along with values for its confidence in those prior values, and then updates the mean and variance estimates with the observed values. The confidence for each of the prior values determines how quickly the model will move from the prior values to the observed values. In the Bayesian updating formulas, the confidence values λ_0 and α_0 give the strength of the prior values in terms of an equivalent number of true observations. Thus, a confidence of 10 indicates that more than 10 observations are necessary before the observed value is given more weight than the prior value, whereas a confidence of 0.1 indicates that even a single observation (or two, for the variance) will overwhelm the prior value.

The expected value $E(i)$ on each unobserved dimension i of the new observation is simply a weighted average of the category means for that dimension. The weights used are the conditional probabilities that the new observation is a member of each of the categories:

$$E(i) = \sum_k P(k|F)\mu_{ik}, \tag{5}$$

where μ_{ik} is the current estimate of the mean of dimension i for category k and is calculated as a weighted average of the prior mean and the mean of the observations of dimension i for category k:

$$\mu_{ik} = \frac{\lambda_0\mu_0 + n_{ik}\bar{x}_{ik}}{\lambda_0 + n_{ik}}. \tag{6}$$

When memory decay is added to the rational categorization algorithm, the primary change is that the past observations are no longer weighted equally. Instead, the weight of each observation starts out as 1 and then decays away. There are a number of possible forms of memory strength decay that might be considered to model the adjustment to category change (cf. Wixted & Ebbesen, 1991). We will limit ourselves to four: exponential and power law decay, each occurring either by item or by time.

Memory decay affects the various ns, as well as the means and variances, $\bar{x}_{ik}$ and s_{ik}^2, of each dimension of each category. The number of observations n is replaced by n_t, the total remaining weight of all past observations at time t. For exponential decay by unit time, n_t is calculated as a function of the decay factor d applicable to the passage of a unit of time:

$$n_t = \sum_{v=1}^{n} d^{\tau_{vt}}, \tag{7}$$

where τ_{vt} is an age function, giving the age in time units of the vth observation at time t. For power law decay by unit time, n_t is calculated as a function of the scaling parameter a for the age in time units and of the power b:

$$n_t = \sum_{v=1}^{n} (1 + a\tau_{vt})^{-b}. \tag{8}$$

For decay by item, the age function τ_{vt} in Equations 7 and 8 is replaced by $n - v$.

Similar formulas yield the total remaining weight of the observations at time t for each category, n_{kt}, and the total remaining weight of the observations for each dimension of each category, n_{ikt}. Similarly, the means and variances, $\bar{x}_{ik}$ and s_{ik}^2, of each dimension of each category are replaced by their weighted values at time t, $\bar{x}_{ikt}$ and s_{ikt}^2. These statistics are calculated as weighted averages of the relevant past observations, using the remaining weight of each observation as its weight for the weighted averages. For example, the mean $\bar{x}_{ikt}$ under power law decay is calculated as follows:

$$\bar{x}_{ikt} = \frac{1}{n_{ikt}} \sum_{v=1}^{n_{ik}} x_{vik}(1 + a\tau_{vt})^{-b}, \tag{9}$$

where x_{vik} is the value of observation v on dimension i for category k.

Because the confidences in the prior values are essentially equivalent to some number of true observations, the addition of strength decay for the observations suggests that the confidences should perhaps decay in an analogous fashion. On the one hand, the prior values can be thought of as distilled information from past observations in other domains. Because this information comes from past observations it may make sense for it to decay in strength. On the other hand, it may be that the process of aggregating information from other domains results in priors that are relatively robust to change and so have no need to decay. To accommodate either of these possibilities, the model allows the confidence in the prior mean to decay in the same way that the observations do λ_{0_t}, except that the decay function for the confidence in the prior mean is fit separately from the decay function for the strengths of the observations. The model's predictions are less sensitive to the confidence in the prior variance α_0, and so this confidence is kept constant to reduce the number of estimated parameters.

In contrast to most symbolic categorization algorithms, including the rational model, exemplar models do not attempt to group their observations into categories. Instead, a similarity measure is calculated between the test stimulus and each past exemplar observation, and this similarity measure is then used to weight the category information contained in the exemplar observations. To model the category-based predictions in these experiments, the same basic procedure was used except that the model used the similarity measure to weight the size information contained in the exemplar observations for the dimensions that need to be predicted. Adapting Nosofsky's (1986; Nosofsky, Clark, & Shin, 1989) notation to the notation already used for the rational model, the similarity s_j between the test stimulus and each past observation j is calculated using the distance d_j between them in terms of the dimensions observed:

$$d_j = \sum_i w_i|F_i - x_{ij}| \tag{10}$$

$$s_j = e^{-cd_j}, \tag{11}$$

where F_i is the ith feature of the test stimulus, x_{ij} is the ith feature of the jth observation, the w_i are attentional weights on the observed dimensions, and c is a sensitivity parameter that scales the absolute distances for calculating the similarity measure.

In the simple exemplar model without decay, the expected value $E(i)$ on each unobserved dimension i of the new observation is the similarity-weighted average of the past-observation values on that dimension.

$$E(i) = \frac{\sum_j s_j x_{ij}}{\sum_j s_j}, \tag{12}$$

where Equation 12 is analogous to Equation 5 for the rational model. In the model that includes decay, the similarity of each past observation is combined with its total remaining weight ω_{jt} at time t:

$$E_t(i) = \frac{\sum_j s_j \omega_{jt} x_{ij}}{\sum_j s_j \omega_{jt}}, \tag{13}$$

where ω_{jt} is calculated with the appropriate exponential or power law decay function, as in Equations 7 and 8 for the rational model (cf. Nosofsky et al., 1992).

In Nosofsky's model, which is used to predict discrete-valued category labels, the similarity-weighted averaging of the category information in the observations is adjusted by response-bias parameters for each of the categories.[3] The model used here adapts these response-bias parameters to serve the case of predicting continuous dimensions rather than discrete category labels. The approach is simply to employ a prior mean of 0.5 (for the 0–1 range) with a strength of λ_{0_t} as in the rational model:

$$E_t(i) = \frac{\lambda_{0_t} 0.5 + \sum_j s_j \omega_{jt} x_{ij}}{\lambda_{0_t} + \sum_j s_j \omega_{jt}}. \qquad (14)$$

As with the rational model, the prior strength is allowed to decay according to its own decay function.

Distinguishing Forms of Memory Strength Decay

The design of Experiment 1 was too simple to yield fine-grained information about the form of the strength decay for past observations. Both the rational and the exemplar models with either exponential or power law decay can produce essentially perfect fits of the performance shown in Figure 2. However, with a more sensitive use of changing categories, it becomes possible to distinguish between predictions based on exponential and power law decay, and between decay by item and by time. In general, exponential decay contrasts with power law decay by having a weaker long-term retention of memories; with power law decay there is a fast initial decay of memory strength but then a much slower decay later on. It is possible, at least in principle, to design sequences of changes that will result in distinctive adjustment patterns for the two decay forms. To distinguish between decay by item and decay by time, participants' performance can be assessed after a delay, during which time has passed but no new observations have been received. Experiment 2 included both a complex change sequence with distinctive adjustment patterns for exponential and power law decay as well as an overnight delay to contrast decay by item and decay by time.

In addition, the design of Experiment 2 allowed a more specific test of the decay function, involving the combined assumptions of power law decay and time decay. If decay occurs by item or if its form is exponential, then the relative strength of all past observations will be the same after a delay. However, if the decay occurs according to a power law and is a function of time, then all observations would decay during the delay with the more recent observations decaying more quickly than the more distant observations. The result of this unequal decay would be a relative reweighting of the observation strength, reducing the relative proportion of strength of the later observations and increasing that of the earlier observations.

An example of the reweighting effect is shown in Figures 3 and 4. Figure 3 shows the strength remaining for 160 observations at three different times, where the strength is determined by the power law decay function $(1 + .189\tau)^{-1.29}$, where τ is the age of the observation. The strength is shown first for Time 161, indicating the strength of all past observations that would apply if a 161st prediction had to be made immediately. The other two times are at 177 and 193, which are, respectively, 16 and 32 time units later. As the figure shows, the strength of the early observations was barely affected by the addition of 32 time units, because those observations were already well past their period of fast decay. However, the most recent observations experienced significant decay from the passage of a few additional time units. In Figure 4, the 160 observations are broken up into five blocks, corresponding to the blocks used in Experiments 1 and 2. Figure 4 shows the proportion of total strength represented by the observations in each block at each of the three times. As time moved from 161 to 177 to 193, the proportion of strength in Block 5 declined from 0.67 to 0.50 to 0.43, while the proportion of strength in the earlier four blocks increased.

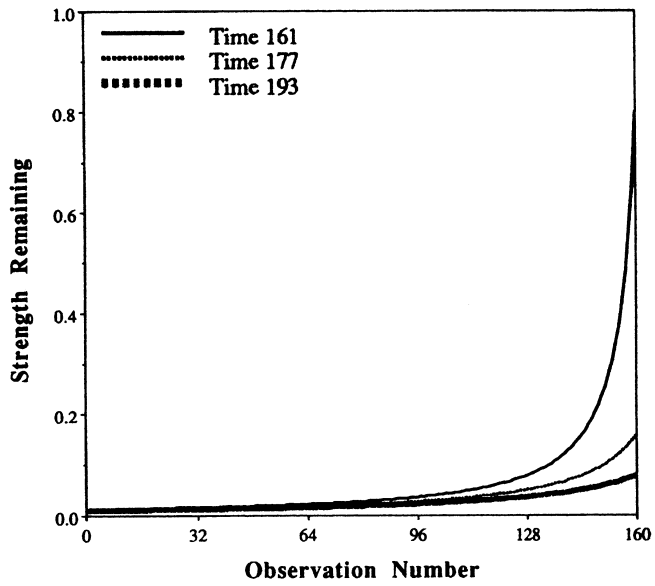

Figure 3. Remaining strength of 160 observations originally observed from Time 1 to Time 160, with strength shown at Time 161, Time 177, and Time 193. Decay occurred according to the power law function $(1 + .189\tau)^{-1.29}$, where τ is the age of the observation.

If participants' memory for the past observations decays according to a power function of time, then their predictions should show a regression toward the equal weighted average after a delay as the more heavily weighted recent observations experience the early fast decay. This prediction of a regression effect is reminiscent of the spontaneous-recovery phenomenon in the paired-associates literature (Brown, 1976; Crowder, 1976).

Experiment 2

Experiment 2 was designed to test how well the categorization models with memory decay could reproduce participant

[3] Using consistent notation, Nosofsky's formula for the probability of responding with category κ becomes

$$P(\kappa) = b_\kappa \sum_{j \in \kappa} s_j \Big/ \sum_k \left(b_k \sum_{j \in k} s_j \right),$$

where the b_k are the response-bias parameters for each category.

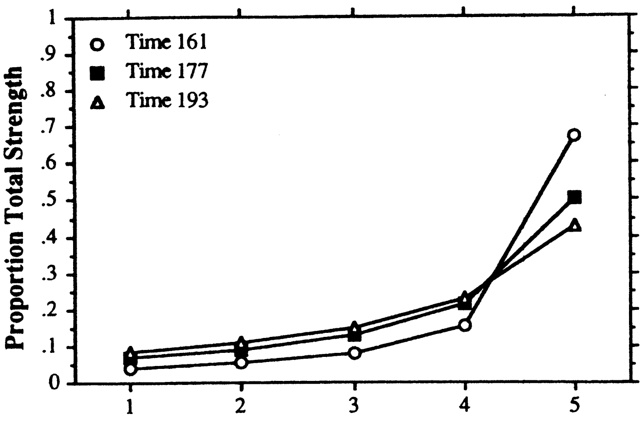

Figure 4. Relative strength remaining at Time 161, Time 177, and Time 193, for the 160 observations of Figure 3, aggregated by blocks of 32 consecutive observations.

performance and to investigate the specific decay function for past observations of changing categories. The experiment included the three different kinds of decay function tests discussed above. To look for adjustment patterns distinctive of either exponential or power law decay, we separated the change on the two changing dimensions to make the overall adjustment pattern more complex. One dimension changed during the first half of the experiment and then remained constant during the second half, whereas the other dimension was constant during the first half and changed during the second half. To contrast item decay and time decay, a 24-hour delay was added for some participants between the first change and the second change. We reasoned that when participants encounter the second changing dimension, they should take extra time to overcome the effects of their past observations of that dimension at its initial constant value. However, if decay occurs by time rather than by item, then the obstacle posed by those past observations should be reduced when the second change occurs after a delay because the strength of the past observations would have decayed. Finally, two tests were added between the two changes to look for the hypothesized regression effect that would result from exemplar reweighting after a delay if memory decay is a power function of time.

Method

Participants. Twenty young adults from the Carnegie Mellon University community participated and were paid $10. None of the participants had been involved in Experiment 1.

Stimuli. The stimuli were pairs of ovals, as in Experiment 1. As before, these pairs of ovals were generated from two different categories that were defined on four dimensions, two of which exchanged values over the course of the learning trials. Unlike in Experiment 1, however, these changes did not occur at the same time. The experiment had two consecutive learning sessions, each of which involved a series of 160 stimuli in five blocks of observations. One of the dimensions exchanged values between the categories in each of the sessions (the EarlyChange and LateChange dimensions, respectively). Figure 5 shows the mean values for the two categories at the beginning and end of the two learning sessions. Going in the forward direction, the heights of the top ovals were exchanged between categories during the first learning session (Blocks 1–5), and the widths of the bottom ovals were exchanged between categories during the second learning session (Blocks 6–10). As in Experiment 1, the direction of the stimulus series was counterbalanced, with half of the participants seeing the series in the forward direction and half seeing it in the reverse direction.

For both changing dimensions, the mean values ranged between 0.2 and 0.8 unit. For the fixed dimensions, Category A had a mean upper width of 0.8 unit and a mean lower height of 0.6 unit whereas Category B had a mean upper width of 0.3 unit and a mean lower height of 0.3 unit. These mean values were used as in Experiment 1 to create a stimulus series for each of the two learning sessions.

As in Experiment 1, the learning sessions required the participant to make a prediction about one of the ovals, which was followed by a display of the correct oval pair along with a superimposed outline of the participant's prediction. The error units and possible comments accompanying the display of the correct pair were the same as before. The one change in the learning procedure was in the size of the anchor oval that replaced the oval to be predicted when each stimulus was initially presented; this oval had a height and width of 0.01 unit, rather than 0.5 unit as in Experiment 1. This modification was made to avoid having the changing dimensions move from being larger than the anchor to being smaller, or vice versa, which might have made the change more obvious.

In addition to the second learning session, two test sessions were added to the experiment. The test sessions were exactly like the learning sessions except that no feedback about the correct oval pair or the success of the participant's predictions was presented. The test sessions occurred between the two learning sessions. The mean values for the stimuli used in these test sessions were the same for each

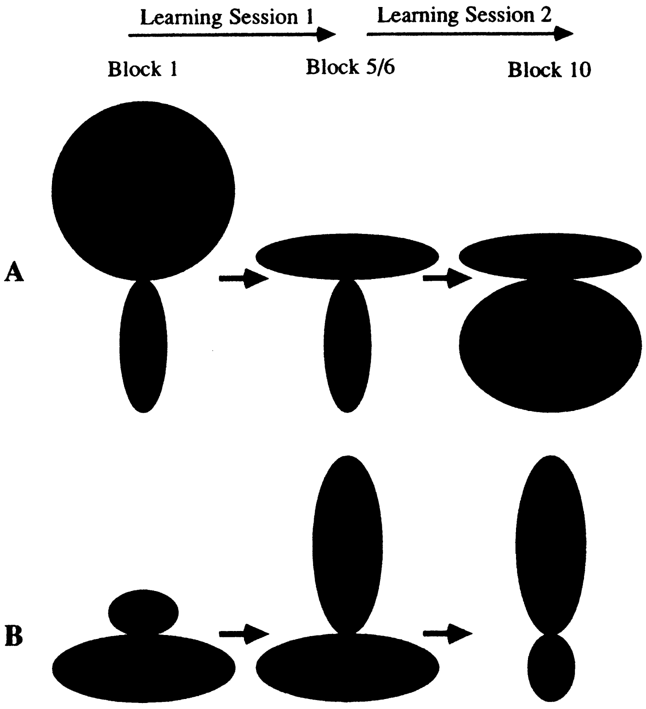

Figure 5. Mean values for stimuli used in Experiment 2, for Categories A and B on Learning Blocks 1, 5, 6, and 10. Stimuli are shown in the forward direction.

category as the values used in Blocks 5 and 6 of the learning sessions. Each of these tests had 16 stimuli, with 8 drawn from each category.

Design. Half of the participants (immediate condition) completed the entire experimental sequence in a single 2-hour session, whereas the other half (delay condition) completed the first learning session and first test on the first day and the second test and the second learning session on the following day. As described above, there was a within-subject contrast between the EarlyChange and LateChange dimensions.

Procedure. The task instructions were the same as before, except for the addition of the second learning session and the two tests. Unlike the previous experiment, only on-line instructions were used. There was no typicality test following the computer sessions.

Results and Discussion

Figure 6 shows the course of learning for all participants over the 10 blocks of the experiment for the two changing dimensions. The results were recoded as in Figure 2 for Experiment 1, except that the two dimensions are shown separately because they change at different times. Note that on the recoded scale, the observations of the EarlyChange dimension are distributed around a mean value of 1.0 for Blocks 5–10, whereas the observations of the LateChange dimension are distributed around a mean value of 0 for Blocks 1–6. The figure collapses over the delay versus immediate timing contrast because there was no significant main effect of timing, $F(1, 18) = 1.104, p > .30, MSE = 0.0366$; and there were no significant interactions of timing with the two changing dimensions, $F(1, 18) = 1.708, p > .20, MSE = 0.0281$; with the 10 learning blocks, $F(9, 162) = 0.615, p > .78, MSE = 0.0232$; or with the two changing dimensions and 10 learning blocks together, $F(9, 162) = 1.299, p > .24, MSE = 0.0120$.

On Blocks 1–5, the EarlyChange dimension replicated the results from Experiment 1. On Block 5, the participants' recoded mean prediction was 0.777 unit, which was significantly greater than the equal weighting average of 0.5 unit, $t(19) = 10.2, p < .001, MSE = 0.0147$. When averaged over Blocks 1–5 of the EarlyChange dimension, the participants' recoded mean prediction was 0.442 unit, which was significantly less than the average true value of 0.5 unit, $t(19) = 2.603, p = .02, MSE = 0.0100$. As in Experiment 1, these results indicate that participants were placing more weight on the most recent observations but that they were giving some weight to the earlier observations as well.

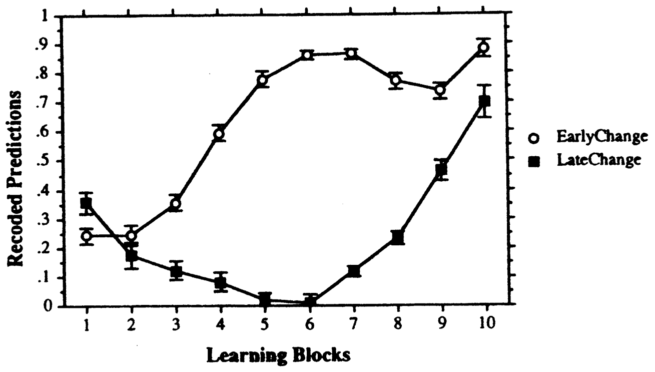

Figure 6. Average predictions on the EarlyChange and LateChange dimensions for Experiment 2. Predictions were averaged over categories, after being recoded so that the true mean of each category of each dimension moved from 0 to 1 over the course of its change. Error bars represent standard errors.

To compare the participants' course of learning with the adjustment patterns of exponential and power law decay requires the category-based prediction models that incorporate those decay functions. The results from the models are discussed in the next section. Here we discuss the two other tests of the form of the decay function.

To test whether decay occurred by item or by time, the adjustment during the first and second learning sessions can be compared for the immediate and delay participants. Although there were no overall effects of timing, any such effects should have occurred only during the second learning session and so would be hard to detect when the 10 blocks from both learning sessions are considered together, as in Figure 6. The two learning sessions are considered separately in Figure 7, which superimposes the five blocks of change for each.

Figure 7 shows that the LateChange dimension was below the EarlyChange dimension for all five blocks of change, for both the immediate and delay participants. The difference between the two dimensions averaged 0.137 over the five learning blocks and was statistically significant, $F(1, 18) = 11.980, p < .01, MSE = 0.0782$. However, this dimension difference did not interact with the timing difference, $F(1, 18) = 0.015, p > .90, MSE = 0.0782$. Thus, the adjustment to the five-block change on the LateChange dimension was held down by the continuing effect of all the Session 1 observations, which had a recoded mean of 0, and this effect was not lessened by inserting a delay between the learning sessions. However, if the decay of past observations was a function of time, then a delay should have weakened the strength of the Session 1 observations on the LateChange dimension, which would have made the LateChange adjustment for the delay participants look more like their EarlyChange adjustment. Because there was no interaction between the dimension and timing differences, this test implies that the memory decay involved in changing categories occurred by item rather than by time.[4]

The other test of the form of the decay function involved looking at the two tests inserted between the learning sessions for evidence that a delay causes the relative weights of the Session 1 observations to shift. This weight shift would increase the relative strength of the earlier observations and decrease the relative strength of the later observations, resulting in a regression toward the earlier values. Such a regression would imply that decay is a power law function of time. The design of Experiment 2 allowed both a between-subjects and a within-subject test for a regression effect. The between-subjects test looked for a main effect of immediate versus delay

[4] The delay participants appeared to learn more successfully on both learning sessions, as reflected in a suggested interaction of the timing difference with the five-block course of learning, $F(4, 72) = 2.023, p = .10, MSE = 0.0149$. Because the immediate and delay participants were treated identically during Session 1, this apparent effect of the timing difference was probably a participant effect.

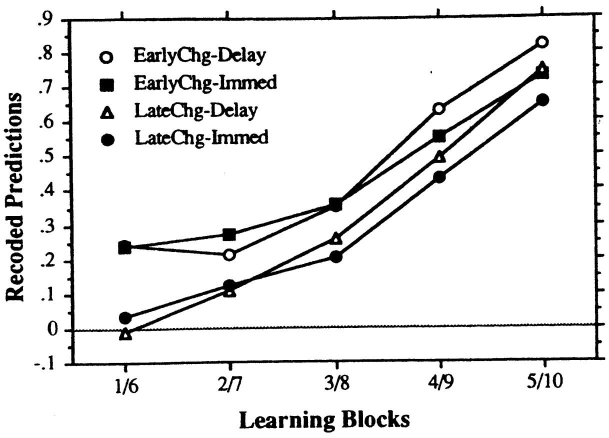

Figure 7. Average predictions over the five-block change for the EarlyChange (EarlyChg) and LateChange (LateChg) dimensions in Experiment 2, shown separately for immediate (Immed) and delay participants. Predictions were averaged over categories, after being recoded so that the true mean of each category of each dimension moved from 0 to 1 over the course of its change.

timing on Test 2. A regression effect would result in lower recoded mean predictions for the delay participants than for the immediate participants on Test 2. However, there was no indication of a between-subjects main effect of timing, $F(1, 18) = 0.012, p > .91, MSE = 0.0320$. The within-subject test looked for a main effect of test number for the delay participants only, whose two tests were separated by a 1-day delay. Although the trend was in the right direction for a regression effect, with recoded mean predictions by the delay participants of 0.783 unit and 0.726 unit on Test 1 and Test 2, respectively, the difference was not significant, $F(1, 9) = 3.723, p = .09, MSE = 0.0044$. This suggestion of a regression effect was investigated further in Experiment 3.

Modeling the Adjustment Pattern of Experiment 2

The rational and exemplar models resulted in a total of eight model variants to be fit to the Experiment 2 adjustment pattern, when both exponential and power law decay were considered for the decay of both the observations and the prior belief about the mean. Because there was no effect of delay in the adjustment pattern of the participants, the simulations assumed that the stimuli were all presented on a single day. The simulations assumed that decay occurred by time, but this was essentially equivalent to decay by item because the categories were equally distributed, and there was no delay between the sessions. The simulations were run on the 320 observations of the two learning sessions, ignoring any effect of the intervening tests. A unit of time was defined as the time required for a participant to process and respond to a single stimulus.

For the simulations using the rational model, the coupling parameter *c* is set to 0.3, the value used in a number of tests of the rational algorithm (Anderson, 1990, 1991). The precise value does not matter, as long as it is in the range for which a small (but nonunitary) number of categories is created. Note from Equations 1–3 that *c* plays a role only in determining the weight on the new category; it does not affect the relative strengths of existing categories, because it cancels out in likelihood ratios of existing categories. In the simulations of Experiment 2, the model had no difficulty grouping around the two clear categories.

The prior mean μ_0 used in the rational model was set at 0.5, the midpoint of the range for each dimension. Anderson and Matessa (1990) suggested that the prior variance σ_0^2 should be set at the square of one-quarter of the stimulus range, allowing values over the entire stimulus range, which would give a value here of 0.0625. The simulations here used a smaller value of 0.018 that was based on the expected number of pieces that each continuous dimension would be broken into to define multiple categories. The intuition for this smaller value is that if one has the prior expectation that a continuous dimension will be divided into small and large values that will be used to define different categories, then one should expect that within each category the values will include only half the entire stimulus range. According to this argument, the prior variance should be set at the square of one-eighth the entire stimulus range, giving a value of about 0.016. The simulations used a slightly larger value because of an involved argument that not all dimensions would need to be divided into pieces for defining categories.

The prior mean confidence, λ_0, was one of the parameters that was fitted to the data. This parameter was critical in modeling the early part of the participants' adjustment path, when they were just learning the category structure and their predictions were biased toward the midpoint of the interval. The model was not particularly sensitive to the prior variance confidence, α_0, which was set at 10. The decay function for the prior weight was applied only to the prior mean confidence; because the prior variance was less critical to this simulation, the confidence in its prior was kept constant.

For the exemplar model, the distances from different dimensions were assumed to be weighted equally, and were each weighted at full strength rather than averaged (which halves the size of the sensitivity parameter *c*). The simplification of equal weighting was used because the model was being fitted to aggregate data that had been counterbalanced across participants. The sensitivity parameter *c* was fit as an additional parameter, because it was crucial in determining whether the model would take information from only the few closest-fitting exemplars or whether it would aggregate relatively equally across a large number of nearby exemplars.

The strength of the prior mean λ_0 for the exemplar model, along with its decay function, was fit to the data in the same way as for the rational model.

The fitting of the models to the data involved choices of the strength of the prior mean, the parameters of the decay functions for the observations and for the prior (the decay factor *d* in the case of exponential decay, and the scaling parameter *a* and the power *b* in the case of power law decay), and the sensitivity parameter in the case of the exemplar models. The number of free parameters ranged from three to six. The models were fit by searching a series of grids of parameter values. The fit of each set of parameter values for each model was measured by the mean square error between the model's predicted course of learning and the participants' actual course of learning, as represented by the 20 data points in Figure 6. The strength decay over 160 observations that resulted from the best-fitting power decay function is shown in Figure 3.

The adjustment paths of the best-fitting models are shown in Figure 8 averaging over the form of the prior decay, which did not have much effect on the fits. The models all did fairly well, showing that the course of adjustment in Experiment 2 over both the EarlyChange and LateChange dimensions could be successfully modeled by incorporating either exponential or power law decay into the two categorization algorithms.

All the models captured some of the dip that occurred in the participants' predictions of the EarlyChange dimension during Blocks 8 and 9. In the models, this dip occurred because of the category confusion resulting from the change in the LateChange dimension. The LateChange dimension was one of the two dimensions available as a cue for predicting the EarlyChange dimension, and as it began to change the models showed some uncertainty about how to categorize the cue ovals with that dimension. This uncertainty resulted in an increased weight placed on the wrong category, which had a recoded value near 0 for the EarlyChange dimension.

Table 1 gives significance tests for comparing the variance of the model predictions around the participant means with the variance of the participant means around the underlying population means over the 20 points. These tests showed whether the models' predictions were significantly different from the participant means, given the overall uncertainty in using those participant means as estimators of the underlying population means. The numerators for the *F* statistics are the *MSE*s in Table 1. The denominator for the *F* statistics was derived from the *MSE* of 0.0181 for a repeated measures analysis of variance (ANOVA) for the 20 participants on the 20 points. This *MSE* gives the variance of the participants around the participant means, and because there were 20 participants for each mean this implies that the variance of the participant means around the underlying population means was 0.000905.

The *F* tests in Table 1 show that all models except for one showed deviations that were significantly different from the participant means. The one model that was not significantly different was the rational model with power decay of the observations and exponential decay of the prior. In addition, the other rational model with power decay of the observations showed a deviation that had only borderline significance. Note also that the exemplar models with power observation decay achieved closer fits than did the corresponding exemplar models with exponential observation decay, although all the exemplar model fits were significantly different from the participants' performance.

To understand what was determining the performance of the models, it is helpful to look at what the models are paying attention to at the end of the experiment. Figure 9 shows the proportion of weight the models placed on the opposite category, on the prior, and on each of the 10 blocks of observations for the correct category. These weight proportions were actually approximations calculated for an average observation. The weight proportions were calculated with the assumption that observations from the opposite category

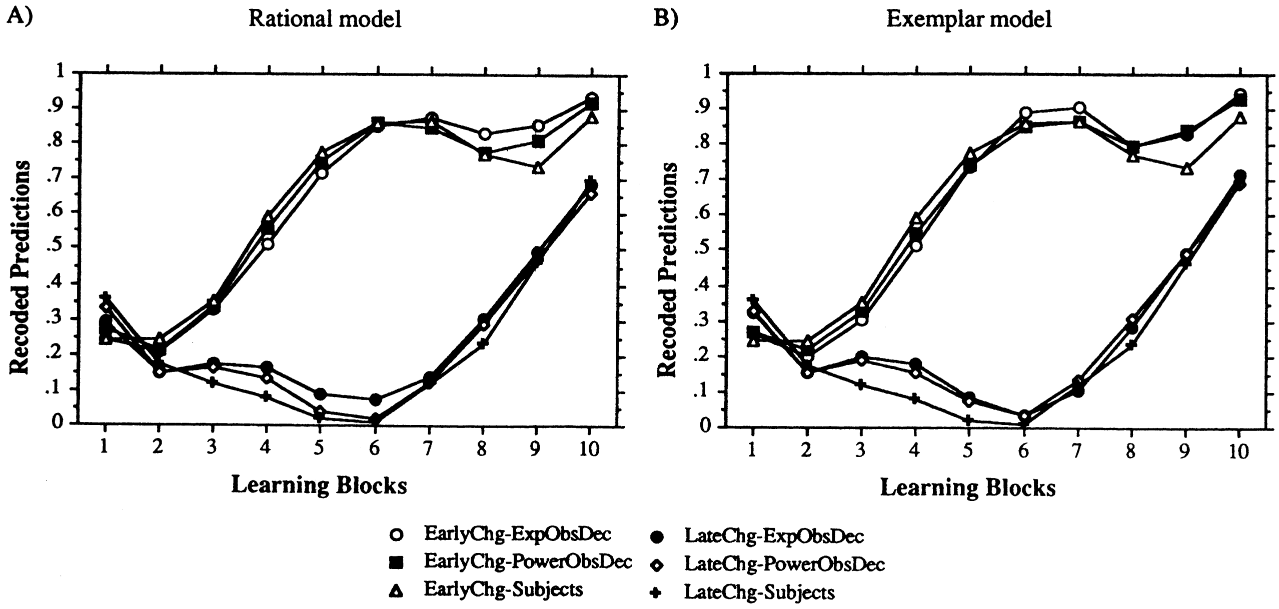

Figure 8. Comparison of model and participant predictions on EarlyChange and LateChange dimensions of Experiment 2 (denoted EarlyChg and LateChg, respectively). Model results are shown separately for power and exponential observation decay (denoted PowerObsDec and ExpObsDec, respectively), averaged over the form of the prior decay. Predictions were averaged over categories, after being recoded so that the true mean of each category of each dimension moved from 0 to 1 over the course of its change. Panel A shows the rational model; Panel B shows the exemplar model.

Table 1
Best Fitting Parameters and Resulting Mean Squared Error (MSE) for Eight Models Fitted to the 20-Block Averages of the EarlyChange and LateChange Dimensions of Experiment 2

Basic model	Obs decay form	Prior decay form	Obs decay parameters	Prior decay parameters	Prior strength	Sensitivity	*MSE*	*F* statistic	*F* test
Rational	Power	Power	$(1 + .239\tau)^{-1.24}$	$(1 + .219\tau)^{-1.69}$	19.5	—	.00159	$F(15, 361) = 1.75$	$p = .04$
Rational	Power	Exp	$(1 + .189\tau)^{-1.29}$	$.961^{\tau}$	4.25	—	.00139	$F(16, 361) = 1.53$	$p = .09$
Rational	Exp	Power	$.971^{\tau}$	$(1 + .0451\tau)^{-.377}$	5.49	—	.00374	$F(16, 361) = 4.13$	$p < .001$
Rational	Exp	Exp	$.971^{\tau}$	$.998^{\tau}$	3.93	—	.00387	$F(17, 361) = 4.28$	$p < .001$
Exemplar	Power	Power	$(1 + .142\tau)^{-1.36}$	$(1 + .187\tau)^{-1.51}$	11.9	.795	.00323	$F(14, 361) = 3.57$	$p < .001$
Exemplar	Power	Exp	$(1 + .226\tau)^{-1.21}$	$.949^{\tau}$	3.70	.750	.00252	$F(15, 361) = 2.78$	$p < .001$
Exemplar	Exp	Power	$.970^{\tau}$	$(1 + .168\tau)^{-1.10}$	11.9	.781	.00392	$F(15, 361) = 4.33$	$p < .001$
Exemplar	Exp	Exp	$.971^{\tau}$	$.963^{\tau}$	6.58[a]	.701[a]	.00331	$F(16, 361) = 3.66$	$p < .001$

Note. Obs = observed; Exp = exponential.
[a]Other parameter values differing in the third digit produced the same *MSE*.

would show a difference of 0.6 on each dimension and that observations from the same category would show a difference of 0.07 from the category mean and of 0.1 from other observations in the same category. These latter differences were derived from the distribution of the stimulus values with a variance of 0.005 around the mean for each dimension of each category. The weight proportions were calculated with the decay that would apply if the models were being run on a 321st observation.

Panel A of Figure 9 shows the weight proportions for the rational models, collapsed over the form of the prior decay. The figure shows that the models with power observation decay focused more attention on the most recent block of observations (Block 10) and more on the earlier blocks of observations (Blocks 1–7) than did the models with exponential observation decay. The net result was that the power models were able to adjust quickly to new information while still being restrained somewhat by the observations from the beginning of the learning series. The rational models with power observation decay put essentially no weight on the prior or the opposite category at the end of the experiment. In contrast, the rational models with exponential decay put a substantial amount of weight on the prior at the end of the experiment. With the exponential models, little weight remained on the observations from the early part of the experiment and so without some restraining influence these models would have overshot the participants during the second part of the experiment. The prior, with a value of 0.5 (on both the original and recoded scales), was used by the exponential models to provide a moderating force against the attention on the most recent observations. This emphasis on the prior at the end of the experiment was a way for the models with exponential observation decay to make up for their inability to maintain some significant strength on the early observations. Note that even with the moderating force of the prior, the rational models with exponential observation decay were slower to adjust to the change on the EarlyChange dimension and then overshot the participants at the end of the experiment (Figure 8, Panel A).

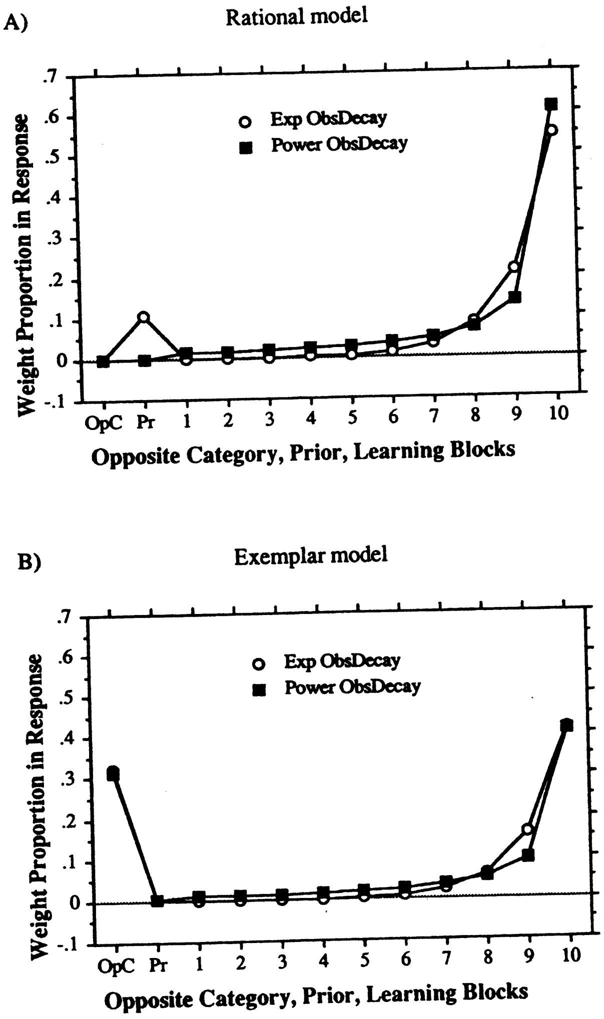

Figure 9. Proportion of weight placed by the models at the end of Experiment 2 on the opposite category (OpC), on the prior (Pr), and on each of the 10 blocks of observations for the correct category. Results are shown separately for power and exponential observation decay (denoted Exp ObsDecay and Power ObsDecay, respectively), averaged over the form of the prior decay. Panel A shows the rational model; Panel B shows the exemplar model.

Panel B of Figure 9 shows the weight proportions of the exemplar models. The high weight that these models placed on the opposite category was in great contrast to the performance of the rational models. For the rational models, a coupling parameter of 0.3 assured that the models would produce two clear categories for these stimuli. However, for the exemplar models, the category discrimination was determined by the size of the sensitivity parameter. Because this parameter was estimated when fitting the models, it was possible to end up with a very fuzzy distinction between the categories. In this case, the exemplar models blurred the categories to provide a moderating force to prevent overshooting the participants. This was possible with these stimuli because both categories moved in tandem for both dimensions, with the category values coming together and then crossing.

In summary, the three tests of the form of memory decay in Experiment 2 were all consistent with power law decay that occurs by item: The overall adjustment pattern was better fit by a model incorporating power law decay than by a model incorporating exponential decay; the adjustment in Session 2 did not seem to be made easier by inserting a day's delay, as would occur if the Session 1 observations weakened over time; and the predictions from the Session 1 observations did not show a significant regression toward earlier values after a delay, as would occur if they weakened according to a power law from the passage of time alone. However, there was a trend in Experiment 2 toward a regression effect, suggesting that there may have been some decay occurring by time as well. In Experiment 3 we investigated the regression effect further, and in Experiment 4 we investigated how well the various decay functions could be fit to a different adjustment pattern.

Experiment 3

Several flaws in the design of the previous experiment made it difficult to interpret the nonsignificant trend toward regression. First, some of the test stimuli gave participants additional observations on the changing dimension at the ending value, which might have inhibited a tendency to show regression. Second, because the change occurred by having the two categories exchange values on the changing dimension, some participants might have become confused when the category values crossed. This confusion might have led these participants to construct new category definitions after the category values crossed, and as a result to show little regression. Third, given the nature of the category change in the previous experiment, any apparent regression could have been caused by a blurring of the category distinction after the delay, rather than by a true regression.

Experiment 3 was designed to overcome the above three weaknesses in our previous attempt to find a regression effect. First, the tests did not use the changing dimensions as stimuli, so no new observations on them were provided. Second, to reduce confusion, the category values on the changing dimensions did not cross but instead maintained the same ordinal relation throughout the learning session. Third, the nature of the changes was altered so that true regression could be distinguished from category blurring. For each participant, there was one dimension on which the categories moved closer together and another on which the categories moved farther apart. For the dimension on which the categories move closer together, a regression would cause the categories to move farther apart, thus working in the opposite direction from category blurring. Finally, new stimuli with four separable dimensions were used rather than the oval pairs to eliminate possible coding problems with the integral height and width dimensions of the ovals in the first two experiments.

Along with these changes to the stimulus structure, a new learning task was added for this experiment. In our previous test of the regression effect, we gave all participants the task of predicting two missing dimensions for each stimulus from the other two dimensions that were given as a cue. We reasoned that if the task were made more like a memory retention task, with more emphasis on remembering the individual observations than on finding a prediction strategy, then we might find a stronger effect of time. This additional task thus served both as a check on the sensitivity of our test for regression and as an opportunity to contrast the effect of time for prediction and memorization encoding. In the memorization version of the learning task, participants were presented the stimuli to memorize in groups of three and then had to reconstruct two dimensions of the stimuli they had just seen when given two dimensions as a cue. The prediction learning task was altered to give comparable stimulus exposure by adding a review after each group of three predictions. The learning sessions for both tasks were followed by prediction tests.

In Experiment 3, we also changed the testing procedure to include instructions directing participants to make their predictions like the beginning or like the ending part of the learning sequence. We added this instruction manipulation to test the extent of participants' explicit knowledge of the change.

Method

Participants and payment. The participants were 64 young adults from the Carnegie Mellon University community, who received $5 or course credit for their participation, plus a bonus of up to $7. None of the participants had been involved in our previous experiments about category change. The bonus was linked to the error-units score associated with each prediction or reconstruction: The participants received a bonus of 2¢ for an error-units score between 8 and 14, a bonus of 4¢ for a score between 4 and 7, and a bonus of 6¢ for a score of 3 or less. The participants received bonuses for their answers in both the learning and the test sections of the experiment, although because the test sections provided no correction, the participants were informed only about their total bonus at the end of each section.

Stimuli. The stimuli were caricatures of teapots, inspired by stimuli used by Ahn and Medin (1992). The stimuli were presented on a Macintosh IIci computer with a monochrome two-page monitor. Each teapot had four relevant dimensions: the base height, the handle thickness, the lid width, and the spout elongation. The stimuli were constructed so that they would fall into two categories. Two of the dimensions were fixed for both categories, whereas the other two dimensions each changed for one of the categories. The two fixed dimensions provided unambiguous information about category membership. The change on the other two dimensions took place over a series of five blocks of observations. Within each block, there were 12 stimuli from each of the two categories, for a total of 120 stimuli for the learning section of the experiment. Figure 10 shows the mean values for the two categories for Blocks 1 and 5 for one of four stimuli conditions (discussed below).

For both changing dimensions, only one of the categories changed while the other stayed constant. In the example shown in Figure 10, the changing dimensions are the lid width and the spout elongation. The width of the lid increased for Category A and was always larger than the constant width of Category B. The elongation of the spout decreased for Category B and was always larger than the constant elongation of Category A. The stimuli for all participants included one dimension for which the categories moved farther apart, like the lid width in Figure 10, and one dimension for which the categories moved closer together, like the spout elongation in Figure 10.

In the example shown in Figure 10, the dimension where the categories moved apart is also the dimension where the changing category was increasing in size, whereas the dimension where the categories moved together is also the dimension where the changing category was decreasing in size. This pairing was counterbalanced, so that the contrast of categories moving together or apart was not confounded with change that was increasing or decreasing in size. In addition, half of the participants saw the changes take place on base height and handle thickness, instead of the lid width and spout elongation as shown in Figure 10.

The four dimensions were each coded on a unit interval, although the meaning of this code was different for each dimension. On the screen, the base height ranged from 2.60 in. (66.0 mm) (coded as 0) to 4.68 in. (118.9 mm) (coded as 1). The handle thickness ranged from 0.06 in. (1.5 mm) (0) to 0.91 in. (23.1 mm) (1). The lid width ranged from 0.13 in. (3.3 mm) (0) to 2.34 in. (59.4 mm) (1). The spout elongation ranged from a half oval with a width diameter of 0.85 in. (21.6 mm) and a height radius of 0.99 in. (25.1 mm) (0) to a half oval with a width diameter of 1.98 in. (50.3 mm) and a height radius of 0.43 in. (10.9 mm) (1).

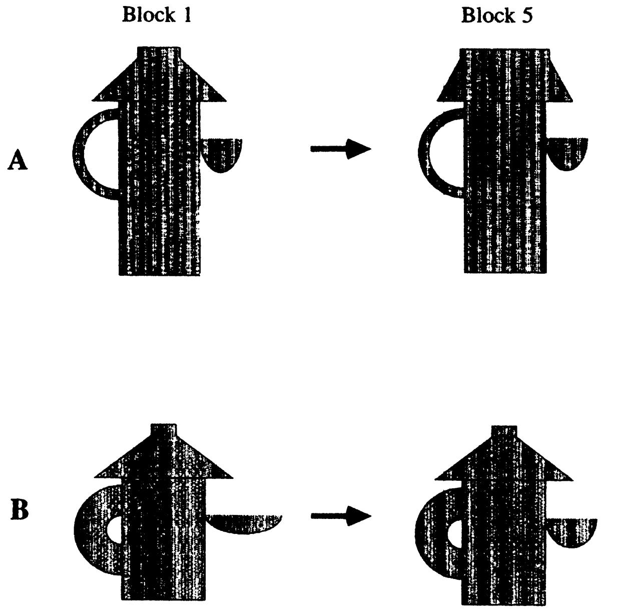

Figure 10. Mean values for stimuli used in Experiment 3, for Categories A and B on Learning Blocks 1 and 5. Stimuli with changes occurring on the lid and spout are shown, with the dimension where categories moved apart paired with increasing size and the dimension where categories moved together paired with decreasing size.

In terms of the unit value coding, the constant dimensions always had a mean value of either 0.2 or 0.8, depending on the category. For the changing dimensions, the value depended on the condition. For the example shown in Figure 10, in which the dimension moving together was the dimension where the changing category decreased in size, the constant categories had a mean value of 0.2, whereas the changing categories moved from 0.9 to 0.4 for the dimension moving together, and from 0.4 to 0.9 for the dimension moving apart. In the counterbalancing condition, in which the dimension moving together was the dimension where the changing category increased in size, the constant categories had a mean value of 0.8, while the changing categories moved from 0.1 to 0.6 for the dimension moving together, and from 0.6 to 0.1 for the dimension moving apart. As in the previous experiments, these mean values were used to create a stimulus series for each of the counterbalancing conditions. The same stimuli were used for all participants within each condition but were randomly ordered within each block for each participant.

As in the previous experiments, the learning sessions in the prediction task condition required the participants to make a prediction about two missing dimensions of each stimulus. Each prediction was followed by a display of the full stimulus along with a superimposed outline of the participant's prediction. The error units and possible comments accompanying the display of the correct pair were the same as in Experiments 1 and 2 with the addition of bonus information on applicable trials. After predictions and corrections of three stimuli, the participants were shown the complete versions of those stimuli again in a random order as a review.

The memorization-task participants had the same two prediction–correction and review sections as the prediction-task participants did, but the order of the sections was reversed. For each group of three stimuli, the memorization-task participants received their review section first, with the instructions to memorize the stimuli to be able to fill in the test items that would immediately follow. After memorizing the three stimuli, the memorization-task participants then saw them in a random order in a memory test, which was identical to the prediction task for the prediction-task participants, including a correction for each reconstructed stimulus.

The reviewing time in the prediction-task condition and the previewing time in the memorization-task condition were set at either 5 s or 15 s for each exemplar and were varied across participants. Participants initiated the viewing of each set of three stimuli, and then the stimuli were presented in series, each for the specified exposure. The viewing time was controlled to ensure equal stimulus exposure for participants in the prediction- and memorization-task conditions. Two times were used to allow for more time-consuming viewing strategies, particularly for the memorization-task participants who might have resorted to simpler prediction methods if the viewing time seemed too short to try to memorize individual stimuli.

In the prediction or reconstruction task, all pairs of dimensions were used equally as stimulus cues within each learning block. The anchor values for the dimensions to be predicted or reconstructed were counterbalanced across participants, with 0 for half of the participants and 1 for half of the participants.

After the learning phase, all participants completed two pairs of prediction tests, one pair immediately after the learning task and one pair following a delay of 24 h. In each pair of tests, one test included instructions to make the test stimuli look like the teapots at the end of the learning sequence, whereas the other test included instructions to make the stimuli look like the teapots at the beginning of the learning sequence. The order of these tests was the same on both days and was counterbalanced across participants. For the prediction tests, only the two constant dimensions were used as cues, and the participants' task was to fill in the two dimensions that had changed during the experiment. There were 24 items in each of the four prediction tests, half corresponding to each of the two categories. The test stimuli were constructed in the same way as the learning task stimuli, so that the values of each dimension for each category were distributed with a variance of 0.005.

Design. There were two between-subjects conditions: the learning task and the time set for the viewing of each stimulus during the review or preview section. There were three within-subject conditions: the contrast between the first and second pair of prediction tests, the contrast between the beginning and ending instructions within each set of prediction tests, and the contrast between the dimensions moving together and moving apart.

Procedure. Participants received on-line instructions that were appropriate to their task. The wording of these instructions for the learning session was as identical as possible for the two tasks. Prediction-task participants were told that their task was to predict the shapes of the teapots and that they would review the teapots in groups of three after their predictions. Memorization-task participants were told that their task was to memorize the teapots in groups of three and that the teapots would then be shown in a test. Both groups of participants were told for the prediction or test section of the learning session that their task was to "fill in the other two parts so that the complete teapot is correct."

As before, the instructions did not mention that the teapots were structured around categories or suggest that the definitions of a correct teapot might change over the course of the experiment. Participants were not informed at the start of the experiment that the learning task would be followed by a prediction test, so for the memorization-task participants the learning session instructions did not suggest that they should prepare for such a test.

Once they were set up on the computer, participants proceeded automatically from the on-line instructions to a practice session and then to the learning task and the first pair of prediction tests. The second session similarly sequenced automatically through the second pair of prediction tests. Participants had unlimited time for all sections of the experiment except for the viewing sections.

Results and Discussion

Figure 11 shows the course of learning for the prediction- and memorization-task participants over the five blocks of the experiment for the category experiencing the change for each of the two changing dimensions. The results are recoded as for the previous experiments so that after recoding the true mean changes from 0 in Block 1 to 1 in Block 5.

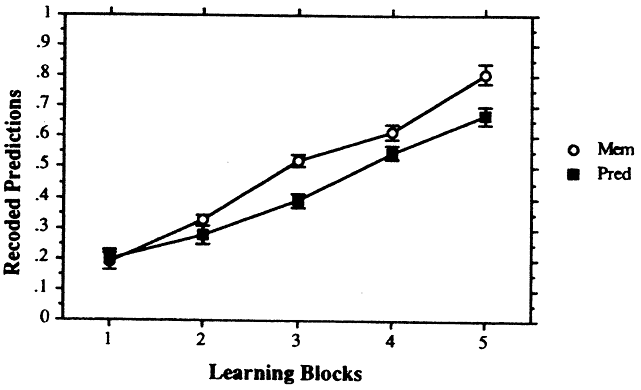

Figure 11. Predictions on the changing categories of the changing dimensions during the learning session of Experiment 3, shown separately for the prediction-task and memorization-task participants (denoted Pred and Mem, respectively). Predictions were averaged over dimensions, after being recoded so that the true mean of the changing category of each changing dimension moved from 0 on Block 1 to 1 on Block 5. Error bars represent standard errors.

The prediction-task participants replicated the adjustment paths of the participants in Experiments 1 and 2. On Block 5, the prediction-task participants' recoded mean prediction was 0.673 unit, which was significantly greater than the equal weighting average of 0.5 unit, $t(31) = 5.897, p < .001, MSE = 0.0275$. Averaging over Blocks 1–5, the prediction-task participants' recoded mean prediction was 0.418 unit, which was significantly less than the average true value of 0.5 unit, $t(31) = 4.824$, $p < .001$, $MSE = 0.00932$. As with the previous experiments, these results indicate that the prediction-task participants were placing more weight on the most recent observations but that they were giving some weight to the earlier observations as well. As would be expected by the difference in learning tasks, the reconstructions by the memorization-task participants followed the change in true values more closely than did the predictions of the prediction-task participants. There was a significant interaction between the learning task and the five-block course of learning, $F(4, 240) = 3.183, p = .01, MSE = 0.0177$.

There were no interesting significant effects of the viewing time on either the course of learning or the prediction tests, and so we do not discuss that manipulation further.

Despite the differences resulting from the two learning tasks that were evident during the learning phase, the two groups of participants performed the same on the first-day tests. The participants averaged 0.708 unit on the first test with ending instructions and 0.459 unit on the first test with beginning instructions. There was no significant effect of task on these tests, $F(1, 60) = 0.385, p > .53, MSE = 0.0398$, and $F(1, 60) = 0.060, p > .80, MSE = 0.0461$, for the first ending test and the first beginning test, respectively.

Between the tests on the first and second days, the participants showed a within-subject regression effect, $F(1, 60) = 37.683, p < .001, MSE = 0.0096$. Averaged over the beginning and ending instructions, the predictions showed a drop of 0.075 unit, moving from 0.583 unit on Day 1 to 0.508 unit on Day 2. The regression effect did not interact significantly with the learning task, $F(1, 60) = 1.965, p = .17, MSE = 0.0096$. However, there was a nonsignificant trend in the predicted direction, with the prediction-task participants showing a regression of 0.058 and the memorization-task participants showing a regression of 0.092. Thus, the trend suggests that regression may be stronger when exemplar encoding is stronger because of memorization instructions.

To contrast the effects of true regression and of category blurring, the experiment included both a changing dimension that moved together and a changing dimension that moved apart. For the change where the categories moved apart, regression would be indistinguishable from category blurring, because both would move the changing category back toward the constant category. In contrast, for the change where the categories moved together, category blurring would still move the changing category toward the constant category, but regression would move the changing category away from the constant category. There was no significant interaction between the regression effect and the type of change (together or apart), $F(1, 60) = 0.567, p > .45, MSE = 0.0329$. This indicates that the regression effect was not an artifact of category blurring.

Participants showed a strong and significant effect of the instructions on their test predictions, $F(1, 60) = 62.075, p < .001$, $MSE = 0.0561$. Averaged over the two days, the predictions for the instructions differed by 0.233, with participants averaging 0.429 and 0.662 for the beginning and ending instructions, respectively. Thus, the participants in aggregate indicated significant knowledge of the nature of the change that had occurred in the category definitions. There was no predicted interaction between the instruction effect and the learning task, and none was found.

There was no statistically significant interaction between the regression and instruction effects, $F(1, 60) = 1.710, p > .19$, $MSE = 0.00957$. Participants regressed from 0.708 to 0.617 for the ending instruction tests and from 0.459 to 0.400 for the beginning instruction tests. This lack of interaction contradicts a possible model of the instruction effect in which the participants respond by selectively using only the earlier or the later observations from the learning session. If the instruction effect were produced by this kind of segregation, then only the ending tests would show regression because the beginning observations would have already gone through their period of fast initial power law decay and so would experience little reweighting overnight.

The regression and instruction effects were also statistically independent across participants, showing only a weak and nonsignificant correlation, $r(62) = .105, p > .40$. This statistical independence implies that the apparent display of joint regression and instruction effects was not an artifact of averaging over participants. It would be plausible to suppose that some participants used a strategy that produced a strong instruction effect and a weak regression effect (say by focusing on updating rules rather than on remembering exemplars), whereas other participants used a strategy that produced a strong regression effect and a weak instruction effect (say by focusing on remembering exemplars rather than on updating rules). Such a strategy difference would have produced both regression and instruction effects in the aggregate, but it would also have produced a negative correlation between the effects across participants, which was not found.

The purpose of Experiment 3 was to provide a cleaner test of the regression effect, whose existence was suggested by Experiment 2. The data show a within-subject regression effect of about 0.075 that was highly statistically significant. The modest size of the regression effect was comparable to the size of the trend found in Experiment 2. This regression effect was not due to category blurring. There was a nonsignificant trend suggesting that the regression effect may have been stronger when exemplar encoding was stronger because of memorization instructions. In addition, the experiment shows that participants had some knowledge of the change that they were able to use to respond differentially to the beginning and ending test instructions. This instruction effect did not interact with the regression effect, and the two effects were statistically independent.

Modeling the Regression Effect

The regression effect seen in Experiments 2 and 3 suggests that some reweighting occurred among the past observations overnight, increasing the relative weight of the earlier observations while decreasing the relative weight of the later observations. Such reweighting could have been produced with power law decay of the observations over the delay. However, there was no indication of a delay effect on the adjustment pattern in Experiment 2, as measured by the speed of adjustment for the different changes shown in Figure 7. The apparent contradiction can be explained as resulting from time decay over 24 h that was small relative to the item decay from 160 new observations. If this was the case, then the impact of new observations in Session 2 would have quickly overwhelmed the difference between the immediate and delay participants that could be seen in test performance before any new observations had been added.

This explanation can be illustrated with the two rational models that caused the observations to decay according to a power law. When the models were run on both the learning and test portions of Experiment 2, it was possible to produce a regression of about 0.057 (the size found in Experiment 2) without greatly affecting the course of adjustment by the LateChange dimension after the delay. The models were run on both the learning and test portions of Experiment 2, using the best fitting parameters shown in Table 2. The learning session portion of the simulations was run exactly as before. To

Table 2

Best Fitting Parameters, Resulting Mean Squared Error (MSE), and Significance Tests for Eight Models Fitted to the 60-Block Averages of the Single-Change and Reverse-Change Dimensions of Experiment 4

Basic model	Obs decay form	Prior decay form	Obs decay parameters	Prior decay parameters	Prior strength	Sensitivity	*MSE*	*F* statistic	*F* test
Rational	Power	Power	$(1 + .0751\tau)^{-1.53}$	$(1 + .00943\tau)^{-1.44}$	8.07	—	.00430	$F(55, 870) = 1.17$	$p = .19$
Rational	Power	Exp	$(1 + .0949\tau)^{-1.41}$	$.993^{\tau}$	6.35	—	.00387	$F(56, 870) = 1.05$	$p = .38$
Rational	Exp	Power	$.975^{\tau}$	$(1 + .100\tau)^{-.175}$	10.4	—	.00625	$F(56, 870) = 1.70$	$p < .01$
Rational	Exp	Exp	$.974^{\tau}$	$.997^{\tau}$	9.11	—	.00544	$F(57, 870) = 1.48$	$p = .01$
Exemplar	Power	Power	$(1 + .0900\tau)^{-1.50}$	$(1 + .00696\tau)^{-2.00}$	3.02	7.51	.00412	$F(54, 870) = 1.12$	$p = .26$
Exemplar	Power	Exp	$(1 + .0710\tau)^{-1.59}$	$.992^{\tau}$	3.10	6.85	.00380	$F(55, 870) = 1.03$	$p = .42$
Exemplar	Exp	Power	$.975^{\tau}$	$(1 + .0100\tau)^{-.959}$	5.41	7.39	.00554	$F(55, 870) = 1.51$	$p = .01$
Exemplar	Exp	Exp	$.977^{\tau}$	$.993^{\tau}$	7.82[a]	4.39[a]	.00540	$F(56, 870) = 1.47$	$p = .02$

Note. Each dimension is separated into start-midpoint and start-endpoint conditions. Obs = observed; Exp = exponential.

[a] Other parameter values differing in the third digit produced the same *MSE*.

generate Test 1 performance, the models were used to make predictions about each of the Test 1 items, without recording any of the test items as observations. During this first test, the observation strengths were fixed at the values attained at the end of Learning Session 1, with no further decay occurring during the test. Test 2 performance was then obtained by decaying the observation strengths by four time units, which was just enough decay to result in a regression of about 0.057 from Test 1. Although there was an effect of this four-unit delay on the LateChange adjustment, the effect was very small. On Block 6, the immediate and delay models differed by only 0.004 and 0.003 on the EarlyChange and LateChange dimensions, respectively. Thus, the models suggest that it was hardly surprising that the delay showed no significant impact on the LateChange adjustment in Experiment 2.

Experiment 4

Experiment 4 was designed to replicate the finding that the adjustment pattern was better fit by a model incorporating power law decay than by a model incorporating exponential decay. As in Experiment 2, this experiment had two dimensions experiencing change, but both the structure of the changes and the nature of the stimuli were different than in Experiment 2. Experiment 4 used the teapot stimuli from Experiment 3, with the basic change on two dimensions defined in the same way as it was in that experiment. This basic change appeared during the middle third of Experiment 4 for both dimensions. During the first third of the experiment both dimensions were constant at the initial value. During the last third of the experiment one dimension stayed constant at the ending value, while the other dimension reversed the change to return to the starting value.

Method

Participants. The participants were 32 young adults from the Carnegie Mellon University community, who received $5 or course credit for their participation, plus a bonus of between $3 and $12. The bonus was linked to the error-units score associated with each prediction, as in Experiment 3. None of the participants had been involved in our previous experiments about category change.

Stimuli. The stimuli were caricatures of teapots, as in Experiment 3. As before, the teapots had four relevant dimensions and were constructed so that they would fall into two categories. Two of the dimensions were fixed for both categories, and the other two dimensions each changed for one of the categories. The two fixed dimensions provided unambiguous information about category membership.

The stimuli were divided into 15 blocks, each with 12 observations from each of the two categories, for a total of 360 stimuli. As in the previous experiment, the change on the two changing dimensions was defined over a series of 5 blocks of observations. In this case, the change for both dimensions occurred on Blocks 6–10. During Blocks 1–5, both of the changing dimensions were constant at the Block 6 value. During Blocks 11–15, one of the changing dimensions (single-change) stayed constant at the value reached on Block 10. The other dimension (reverse-change) reversed the change process during Blocks 11–15, repeating the Block 10 value on Block 11 and then working backward to the Block 6 value on Block 15. The learning sequence was broken up into three consecutive sessions of 5 blocks, comparable to the two learning sessions of Experiment 2 but without tests between sessions.

The mean values for the two categories for Blocks 6–10 corresponded to those used for Blocks 1–5 in Experiment 3. Figure 10 for Experiment 3 shows these values for one of the four stimuli conditions. As in Experiment 3, the stimuli were counterbalanced for (a) both possible pairings of the dimensions moving together or apart with the actual direction of the change (increasing or decreasing), (b) the different pairs of dimensions experiencing the change (handle and base, or lid and spout), and (c) the location of the anchor values on the dimensions that the participants needed to adjust (0 or 1).

As before, the learning sessions required the participants to make a prediction about two of the dimensions of each stimulus, which was followed by a display of the stimulus along with a superimposed outline of the participant's prediction. Like Experiment 3 and unlike Experiment 2, there was no pairing of the dimensions to be predicted, so that every possible pairing of two target dimensions was encountered equally often. Like Experiment 2 and unlike Experiment 3, only a simple prediction learning task was used, with no reviewing session for the stimuli.

Design. There were two within-subject conditions. The first was the contrast between the single-change and reverse-change dimensions. The second within-subject condition was a contrast in the recoded midpoint of the full interval, which was produced by the counterbalancing. For the changes that went from 0.4 to 0.9 and from 0.6 to 0.1 on the full interval, the midpoint of 0.5 on the full interval became 0.2 when the values were recoded as in the previous experiments so that the starting point of the change was 0 and the ending point was 1. For the changes that went from 0.9 to 0.4 and from 0.1 to 0.6 on the full interval, the midpoint of 0.5 on the full interval became 0.8 when the values were recoded. Thus, on the recoded scale there was a contrast between changes with a recoded midpoint value of 0.2 and those with a recoded midpoint value of 0.8. The counterbalancing also produced a between-subjects contrast in how these recoded midpoint values of 0.2 and 0.8 were paired with the single-change and reverse-change dimensions.

Procedure. The overall procedure was similar to that in the previous experiments. The 15 blocks of the learning section were divided into three sections of 5 blocks each. The participants had unlimited time for all sections.

Results and Discussion

Figure 12 shows the course of learning by all participants over the 15 blocks of the learning sequence for the single-change and reverse-change dimensions, respectively. The results were recoded as in the previous experiments. Overall, the participants replicated the previous experiments in showing a pattern of predictions that appeared to emphasize more recent observations.

The results in Figure 12 are shown separately for the two recoded midpoint values. There was a significant interaction between the recoded midpoint contrast and the 15 learning blocks for both the single-change and reverse-change dimensions, $F(14, 420) = 4.015, p < .001, MSE = 0.0573$, and $F(14, 420) = 5.167, p < .001, MSE = 0.0343$, respectively. The figure shows that participants began with recoded predictions in Block 1 that were near the recoded midpoint values. Given these initial starting points, it is not surprising that the participants' predictions approached the true recoded value of zero more quickly when the recoded midpoint was 0.2 than when it was 0.8.

The performance shown in Figure 12 was modeled in the same way as was the performance for Experiment 2. The models yielded separate estimates for the two different recoded midpoint conditions for both the single-change and

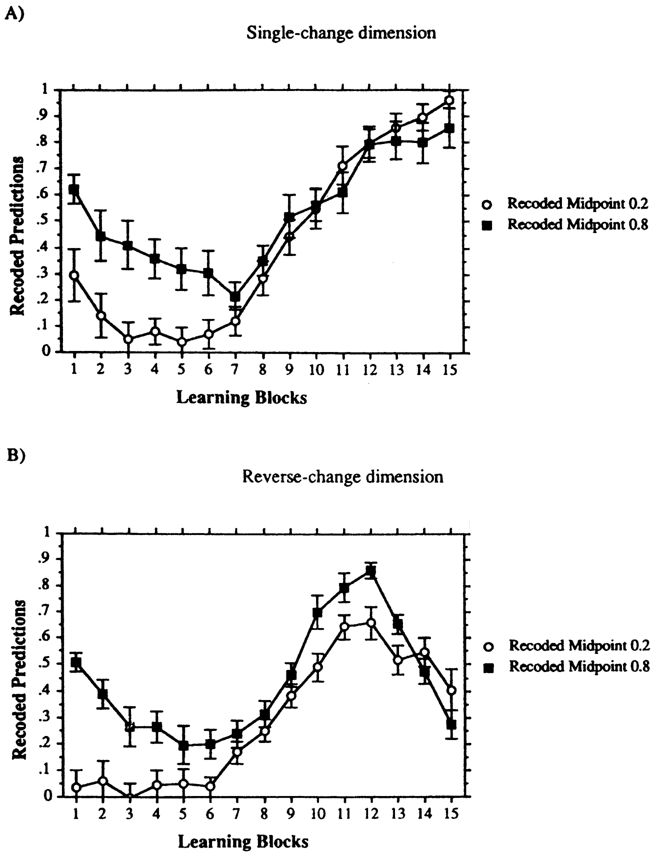

Figure 12. Average predictions for Experiment 4, shown separately for the two recoded midpoints. Panel A shows the single-change dimension; Panel B shows the reverse-change dimension. Predictions were averaged over dimensions, after being recoded so that the true mean of the changing category of each changing dimension moved from 0 on Block 6 to 1 on Block 10. Error bars represent standard errors.

reverse-change dimensions, resulting in comparisons across 60 points between the participants and each of the models. Both the rational and exemplar models were used. As before, four versions of each model were estimated, allowing the decay functions for both the observations and the prior to be either exponential or power law. All versions of the rational model successfully separated the stimuli into two categories. The mean squared errors for the best-fitting models are reported in Table 2.

Table 2 gives significance tests for comparing the variance of the model predictions around the participant means with the variance of the participant means around the underlying population means. The test statistics were calculated as for Experiment 2. A repeated measures ANOVA with one between-subjects variable yielded an *MSE* of 0.0587. With 16 participants for each mean, this *MSE* implied a variance of 0.00367 for the participant means around the underlying population means. When the observations were decayed according to a power law, the model predictions could not be rejected as being significantly different from the participant means, given the overall uncertainty in the estimates of those means. In contrast, when the observations were decayed exponentially, the models could be rejected as showing a significant deviation from the participant means. There was no apparent effect of the form of the prior decay function and no apparent difference between the rational and exemplar models.

Figure 13 shows a comparison of the performance of the participants and the models on each of the four dimension-by-recoded-midpoint conditions. To focus attention on the systematic effects of the observation decay function, the panels of the figure average over the four models for each of the observation decay conditions. These panels show that the large systematic deviations between the power and exponential models came toward the end of the adjustment path for both dimensions and occurred only when the recoded midpoint was 0.8 (Panels B and D). In both cases, the power models were close to the participants and the exponential models were floating above. This contrast is elaborated in Figure 14, which shows the difference between the two recoded midpoint conditions averaged over both dimensions. During the first third of the experiment, participants showed much higher recoded predictions when the recoded midpoint was 0.8 than when it was 0.2, but this difference between the two recoded midpoints disappeared for the participants by the last third of the experiment. By placing weight on the prior and decaying that weight over time, the models had no trouble reproducing the large recoded midpoint difference during the first third of the experiment and decreasing that difference over the last two-thirds of the experiment. However, the models did not fully eliminate the recoded midpoint difference by the last third of the experiment. This difficulty was more pronounced for the models with exponential observation decay.

Figure 15 shows the proportion of weight the models gave to the observations and to the prior at the end of the experiment. The weight proportions given to the prior in this figure explain the substantial recoded midpoint differences that the models showed for the last third of the experiment, as shown in Figure 14. The models used the midpoint of the full interval as their prior. As a result, on the recoded scale there was a 0.6 difference in the prior values used for the two different recoded midpoint conditions. The 0.16 proportion of weight that the exponential models placed on the prior at the end of the experiment explained the 0.11 recoded midpoint difference that they showed during the last third. Similarly, the 0.08 proportion of weight that the power models placed on the prior explained their 0.06 recoded midpoint difference during the last third. The models were using the prior not only to produce the recoded midpoint difference shown by the participants during the first two-thirds of the experiment but also as a moderating force to prevent them from overshooting the participants by adjusting too quickly to new values later in the experiment. Overshooting was more of a problem for the exponential models, because the shape of the decay function made it more difficult to place weight on the earlier observations. In contrast, the power models placed relatively more weight on the earlier observations, and so benefitted from their moderating force on adjustment. As a result, the power models required less moderation from the prior and so produced a smaller recoded midpoint difference during the last third of the experiment. Thus, their performance was closer to that of the participants.

Experiment 4 was conducted to replicate the findings from

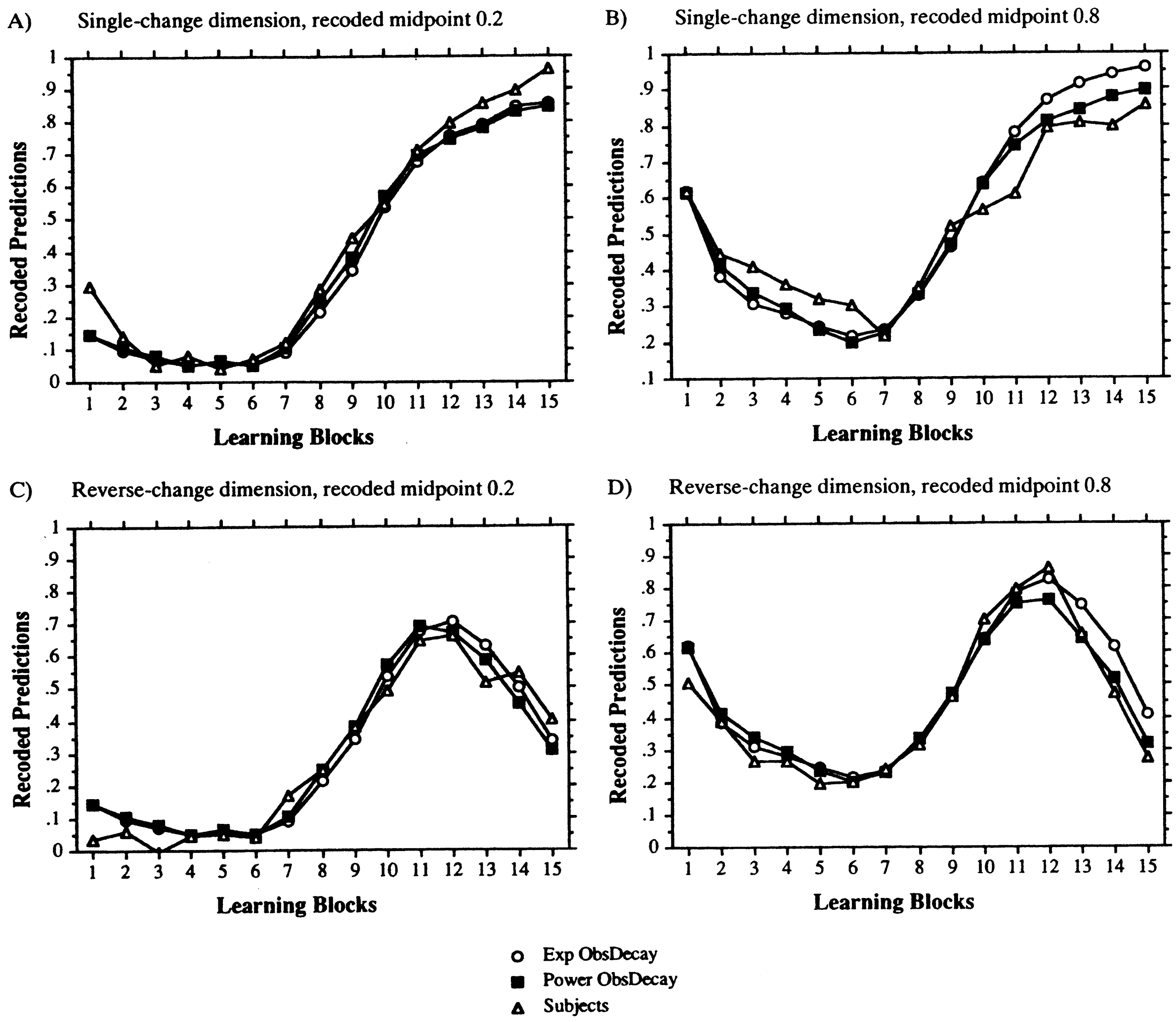

Figure 13. Comparison of model and participant predictions on the four dimension-by-recoded-midpoint conditions of Experiment 4. Model results are shown separately for power and exponential observation decay (denoted Power ObsDecay and Exp ObsDecay, respectively), averaged over the form of the prior decay and the type of model. Predictions were averaged over dimensions, after being recoded so that the true mean of the changing category of each changing dimension moved from 0 on Block 6 to 1 on Block 10. Panel A shows the single-change dimension with a recoded midpoint of 0.2; Panel B shows the single-change dimension with a recoded midpoint of 0.8. Panel C shows the reverse-change dimension with a recoded midpoint of 0.2; Panel D shows the reverse-change dimension with a recoded midpoint of 0.8.

Experiment 2 that the models with power decay of the observations gave closer fits to the participants' adjustment paths than did models with exponential decay of the observations. This experiment found clear preference for power decay of the observations. The models using exponential decay of the observations apparently performed worse because they required a relatively strong weight on the prior to prevent them from overshooting the participants. Unlike Experiment 2, there was no apparent difference between the rational and exemplar models. In Experiment 2, the exemplar models moderated their adjustment by placing substantial weight on the observations from the opposite category. However, the exemplar models could not exploit this option in Experiment 4 because half the time the opposite category would have made their responses more extreme. As a result, the exemplar models for this experiment gave better fits with high sensitivity parameters that clearly differentiated the categories and thus produced fits that were very similar to those of the rational models.

General Discussion

All four experiments reported here found that participants were relatively successful in adapting to category change. This adaptation could be modeled by incorporating memory decay into either the rational categorization algorithm or an exemplar categorization algorithm. In contrast to a model of adaptation that bases performance on the more recent obser-

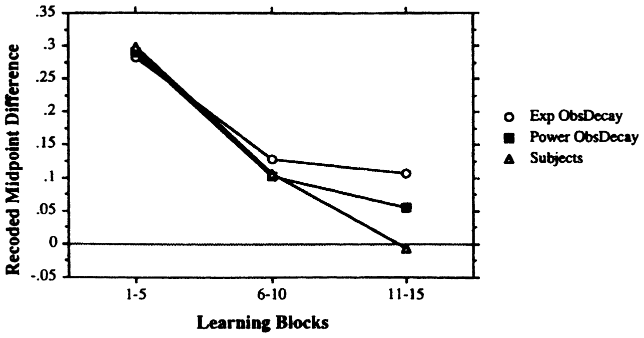

Figure 14. Prediction differences between the recoded midpoint conditions for the models and participants on Experiment 4, averaged over dimensions. Model results are shown separately for power and exponential observation decay (denoted Exp ObsDecay and Power ObsDecay, respectively), averaged over the form of the prior decay and the type of model. Differences were calculated after recoding so that the true mean of the changing category of each changing dimension moved from 0 on Block 6 to 1 on Block 10.

vations alone, a memory decay model assumes that after each stimulus is observed it will have a continued, though diminishing, impact on the remainder of the experiment. A bias toward more recent observations combined with a continuing impact of past observations was seen in all experiments, most strikingly in the later sessions of Experiments 2 and 4.

The specific form of the decay of past observations was closer to a power law function than to an exponential function. In the adjustment to complex change in Experiments 2 and 4, power law decay of observations consistently yielded closer fits to average participant performance than did exponential decay, although the difference was only marginal for the rational models in Experiment 2. The regression effect found in Experiments 2 and 3 strongly supported power law decay of the observations over exponential decay. The experiments showed evidence of both time and item decay, with the time decay resulting from an overnight delay being equivalent to the item decay resulting from about four observations. Overall, the findings were consistent with a rational analysis suggesting that the strength decay in categorization should have the same functional form as that found in the memory retention literature.

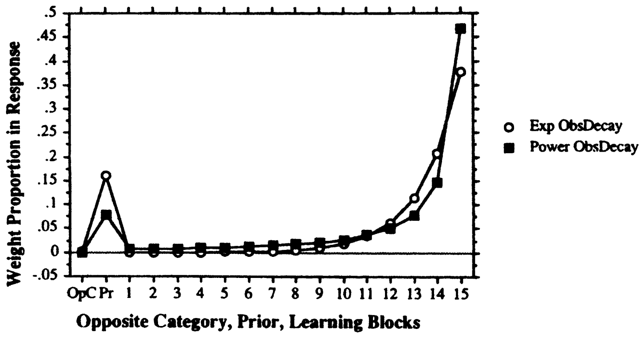

Figure 15. Proportion of weight placed by the models at the end of Experiment 4 on the opposite category (OpC), on the prior (Pr), and on each of the 15 blocks of observations for the correct category. Model results are shown separately for power and exponential observation decay (denoted Exp ObsDecay and Power ObsDecay, respectively), averaged over the form of the prior decay and the type of model.

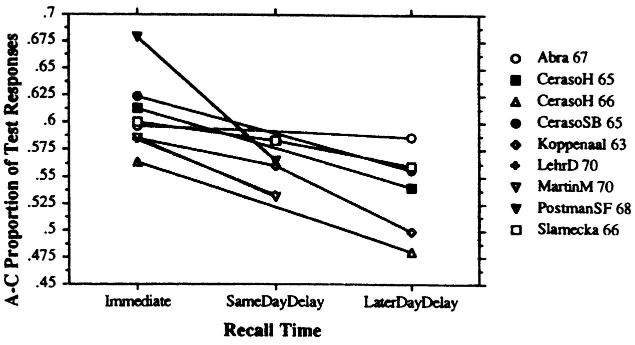

Figure 16. Results of nine paired-associate studies with participants showing spontaneous recovery: Abra (1967), p. 642 (denoted Abra 67); Ceraso & Henderson (1965), p. 301 (denoted CerasoH 65); Ceraso & Henderson (1966), p. 315 (denoted CerasoH 66); Ceraso, Schiffman, & Becker (1965), pp. 262–263 (denoted CerasoSB 65); Koppenaal (1963), p. 313 (denoted Koppenaal 63); Lehr & Duncan (1970), p. 109 (denoted LehrD 70); Martin & Mackay (1970), p. 316 (denoted MartinM 70); Postman, Stark, & Fraser (1968), p. 683 (denoted PostmanSF 68); Slamecka (1966), p. 206 (denoted Slamecka 66). Test results were coded as the proportion of A–C responses.

The regression effect is reminiscent of the spontaneous recovery phenomenon found in the paired-associates literature (see Brown, 1976; Crowder, 1976). That phenomenon was obtained when participants were exposed to an A–B, A–C training sequence and then tested for recall of both lists using the A cues, both immediately after training and then at a delay. On the immediate test, participants gave a higher proportion of associates from the more recent A–C list, but this was followed by an increase in the proportion of associates from the earlier A–B list in the delay test. Figure 16 shows the test results of nine different paired-associates studies in which participants exhibited a spontaneous recovery. The test results were coded in terms of the proportion of A–C responses, so they are comparable to the presentation of the changing category experiments reported here.[5] The delay for the same-day delay tests ranged from 14 min to 6 h, whereas the delay for the later-day delay tests ranged from 1 to 7 days. A simple average over the studies suggests that there was a drop in the proportion of A–C responses of about 0.05 between an immediate test and a same-day delayed test, followed by an

[5] These studies were drawn from a longer list cited by Brown (1976) and were chosen because their results were reported in a way that allows them to be recoded in terms of the proportion of A–C responses. Only one recodable study was excluded from the graph (Goggin, 1966, Lists 1 and 3)—it was excluded because of its unusually low proportion of A–C responses: 0.46 immediately and 0.20 on a same-day delay test (estimated from a published graph). The delay was only 34 s.

additional drop of about 0.02 between a same-day delayed test and a delayed test on a later day. The relatively modest size of spontaneous recovery was comparable to the modest size of the regression effect found for predictions from changing categories.

The finding that memory decay in categorization is closer to a power law than to an exponential gives suggestive evidence that predictions from changing categories are based on exemplar memories. In principle, power law decay can be approximated with several summary statistics rather than with individual exemplar traces, but the similarity to the memory retention literature is striking, and it would be parsimonious to conclude that the same system, or a comparable one, is at work. However, the instruction effect in Experiment 3 shows that participants also had access to knowledge about the nature of the changes that was independent of the decay process producing the regression effect. This finding suggests that participants used generalizations of the stimuli that could reveal the nature of the change, along with exemplars that could produce the regression effect.

Overall, the pattern of adjustment to changing categories in these experiments was qualitatively similar to the adjustment to real-world changing categories, despite the fact that changes in the real world typically take place over days or years, rather than minutes. In both the real world and the laboratory, it appears that people are quite successful at adjusting to category changes but that it takes some time for the new observations to overcome the lingering effect of old observations. The experiments further suggest that there is a cumulative effect of past observations, reflected in the power law form of decay, so that the adjustment process takes longer as one accumulates an increasing number of past observations. It would be interesting to know whether this lingering effect holds over longer times with more observations and whether there is a comparable slowdown in adjustment to change in real-world settings as people acquire more observations in a particular domain. In addition, it would be interesting to know how the effect responds to differences in the abruptness of the change and to participants' expectations about the possibility of change.

References

Abra, J. C. (1967). Time changes in the strength of forward and backward associations. *Journal of Verbal Learning and Verbal Behavior, 6,* 640–645.

Ahn, W. -K., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science, 16,* 81–121.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409–429.

Anderson, J. R., & Matessa, M. (1990). A rational analysis of categorization. In *Proceedings of the Seventh International Machine Learning Conference* (pp. 76–84). Palo Alto, CA: Morgan Kaufmann.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2,* 396–408.

Barsalou, L. W. (1989). Intraconcept similarity and its implications for interconcept similarity. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 76–121). Cambridge, England: Cambridge University Press.

Berg, E. A. (1948). A simple objective technique for measuring flexibility in thinking. *Journal of General Psychology, 39,* 15–22.

Brown, A. S. (1976). Spontaneous recovery in human learning. *Psychological Bulletin, 83,* 321–338.

Busemeyer, J. R., & Myung, I. J. (1988). A new method for investigating prototype learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 3–11.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Ceraso, J., & Henderson, A. (1965). Unavailability and associative loss in RI and PI. *Journal of Experimental Psychology, 70,* 300–303.

Ceraso, J., & Henderson, A. (1966). Unavailability and associative loss in RI and PI: Second try. *Journal of Experimental Psychology, 72,* 314–316.

Ceraso, J., Schiffman, D., & Becker, B. (1965). Recall interference in retroactive inhibition. *Journal of Psychology, 59,* 259–265.

Crowder, R. G. (1976). *Principles of learning and memory.* Hillsdale, NJ: Erlbaum.

Elliott, S. W. (1991). *Steps towards a psychological calculus for game theory.* Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology, 18,* 500–549.

Estes, W. K. (1989). Early and late memory processing in models for category learning. In C. Izawa (Ed.), *Current issues in cognitive processes: The Tulane Flowerree symposium on cognition* (pp. 11–24). Hillsdale, NJ: Erlbaum.

Estes, W. K. (1994). *Classification and cognition.* Oxford, England: Oxford University Press.

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

Goggin, J. (1966). Retroactive and proactive inhibition in the short-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior, 5,* 526–535.

Grant, D. A., & Berg, E. A. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of Experimental Psychology, 38,* 404–411.

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

Keil, F. C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Koppenaal, R. J. (1963). Time changes in strengths of A–B, A–C lists; spontaneous recovery? *Journal of Verbal Learning and Verbal Behavior, 2,* 310–319.

Lehr, D. J., & Duncan, C. P. (1970). Effects of priming on spontaneous recovery of verbal lists. *Journal of Verbal Learning and Verbal Behavior, 9,* 106–110.

Martin, E., & Mackay, S. A. (1970). A test of the list-differentiation hypothesis. *American Journal of Psychology, 83,* 311–321.

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

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 282–304.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 211–233.

Postman, L., Stark, K., & Fraser, J. (1968). Temporal changes in interference. *Journal of Verbal Learning and Verbal Behavior, 7,* 672–694.

Robinson, A. L., Heaton, R. K., Lehman, R. A. W., & Stilson, D. W. (1980). The utility of the Wisconsin Card Sorting Test in detecting and localizing frontal lobe lesions. *Journal of Consulting and Clinical Psychology, 48,* 605–614.

Ruffner, J. W., & Muchinsky, P. M. (1978). The influence of shifting cue validity distributions and group discussion feedback on multiple cue probability learning. *Organizational Behavior and Human Performance, 21,* 189–208.

Rumelhart, D. E. (1989). The architecture of the mind: A connectionist approach. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp. 133–159). Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

Slamecka, N. J. (1966). A search for spontaneous recovery of verbal association. *Journal of Verbal Learning and Verbal Behavior, 5,* 205–207.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Wickelgren, W. A. (1974). Single-trace fragility theory of memory dynamics. *Memory & Cognition, 2,* 775–780.

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science, 2,* 409–415.

Received July 6, 1993
Revision received August 22, 1994
Accepted August 29, 1994 ■