Memory & Cognition
2005, 33 (5), 840-851

Is memory for stimulus magnitude Bayesian?

KEVIN M. SAILOR and MIRIAM ANTOINE
Lehman College, City University of New York, Bronx, New York

This study was designed to determine whether memory for stimulus values is a Bayesian weighting of the magnitude of a stimulus and the central tendency of an exemplar's category (Huttenlocher, Hedges, & Vevea, 2000). In five experiments, participants reproduced the remembered size of a geometric figure drawn from one of two categories whose means for size differed. Reproductions were biased toward the mean of the combined distribution rather than the mean of either category. Reproductions were also influenced by the size of the stimulus on the preceding trial. Neither of these results is entirely consistent with the view that recollections are partially constructed from a consideration of the long-run probabilities established by category membership.

Experience tells us that significant discrepancies can exist between our memories and actual events. For example, many of us have had the experience of walking to our cars only to find that we did not park where we thought we had. A common idea is that these errors are not random but are in fact a result of systematic distortions (Bartlett, 1932). In many cases, these distortions occur because we mistakenly believe that a particular event "followed the norm." Thus, we may walk to the wrong part of a parking lot because we typically park there, even though we did not park there today.

Recently, Huttenlocher and her colleagues (Huttenlocher & Hedges, 1992; Huttenlocher, Hedges, & Duncan, 1991; Huttenlocher, Hedges, & Vevea, 2000) have advanced an explanation of this phenomenon. According to their *category adjustment model*, we utilize two sources of information to reconstruct our memories for particular events or objects. One source is knowledge about the value of the stimulus that we encountered. They argue that instead of thinking about this source of knowledge as exact, we should think of it as a distribution of values centered near the actual value that was encountered. The second source is knowledge about the central tendency of the category of events or objects from which this stimulus was sampled.

The claim of this theory is that we make use of categorical information because it generally improves the accuracy of our memories. According to the theory, factors such as time or interfering events decrease the reliability of our memories because they increase the spread of the values around the actual value of a previously encountered stimulus. Under these conditions, we can improve our accuracy by weighting our inexact representation of an individual item by the mean of the category to which it belongs because we are more likely to encounter stimuli with values close to the mean of the category than stimuli with relatively rare or extreme values. Within the framework of Bayes's theorem, this knowledge about the central tendency of a category provides information about the prior odds of a stimulus value. For example, if I saw a tall individual, my memory for the height of this individual might be represented by a range of values from 6 ft 1 in. to 6 ft 5 in. Although it is possible that I did in fact see a 6 ft 5 in. individual, it is much more likely that I saw someone whose height was closer to 6 ft 1 in. because more people are 6 ft 1 in. than 6 ft 5 in. This difference in the prior probability of the two heights means that the accuracy of my memory could be improved over the long run if I moved my estimate closer to 6 ft 1 in. because I am simply much more likely to encounter 6 ft 1 in. individuals in the first place.

An important implication of this theory is that memory should be biased toward typical or average values. According to the model, the integration of memory for the stimulus and the mean of the category is expressed by the following equation:

$$R = \lambda M + (1 - \lambda)p,$$

where R is the response to a stimulus, M is a random variable that represents a set of values in memory for the stimulus, p is the central tendency of the category, and λ is a weighting parameter that ranges from 1 to 0. Memory for the specific episode will be weighted to a greater extent as the uncertainty of memory for the value of the stimulus decreases and the uncertainty of the category prototype as a predictor increases. Thus, responses to a stimulus should be consistently biased toward the category prototype unless the representation of a stimulus in memory is extremely precise. This pattern of bias translates into reproductions of low-magnitude stimuli that are overestimated relative to their veridical values and re-

Preparation of this article was supported by NIH Grant GM08225. We thank Doug Wedell for comments on an earlier draft of the manuscript. Correspondence concerning this article should be addressed to K. M. Sailor, Psychology Department, Lehman College, 250 Bedford Park Blvd. West, Bronx, NY 10468 (e-mail: kevin.sailor@lehman.cuny.edu).

productions of high-magnitude stimuli that are underestimated relative to their veridical values.

The observation that memory for stimulus magnitude is frequently less extreme than the direct perception of these magnitudes has a long history in the judgment literature and is well established for a variety of tasks. For example, the psychophysical function derived from magnitude estimates of remembered stimuli is frequently flatter than the corresponding psychophysical function for estimates based on direct perception (Kerst & Howard, 1978; Moyer, Bradley, Sorensen, Whiting, & Mansfield, 1978). Similarly, in successive comparative judgment tasks, participants typically overestimate the magnitude of a low-magnitude stimulus relative to that of a second stimulus (i.e., a positive time-order error) but underestimate the magnitude of a high-magnitude stimulus relative to that of a second stimulus (i.e., a negative time-order error) (Hollingworth, 1910). Moreover, this effect increases as the interstimulus interval (ISI) increases (Needham, 1935), presumably because the memory requirements increase as ISI increases. Although a variety of theoretical treatments has been offered to explain these and related phenomena (Estes, 1997; Hellström, 1985; Helson, 1964; Kerst & Howard, 1978; Poulton, 1979), the category adjustment model is unique in claiming that regression to the mean can be characterized as a Bayesian use of category-level information.[1]

Unfortunately, results from a series of psychophysical studies suggest the possibility that a Bayesian process in which the category mean is used to calibrate judgments may not produce this phenomenon. In these studies, stimuli that have been drawn from two different distributions are presented in the same experimental setting. To the extent that these studies demonstrate a contextual effect of the different distributions, they appear to demonstrate contrast rather than assimilation effects (Marks, 1988, 1992; Schneider & Parker, 1990; Wedell, 1995). For example, Marks (1988) found that magnitude estimates of the loudness of a 2500-Hz tone were greater than those of the loudness of a 500-Hz tone of the same amplitude when the set of 2500-Hz tones was generally softer than the set of 500-Hz tones. Wedell (1995) obtained similar results in a comparative judgment paradigm. He asked participants to choose the larger of two successively presented red and blue squares that were separated by a 2-sec interval. Participants reliably preferred the member of the smaller series when the two squares were of equal size. These results in particular suggest that the participants did not follow Bayesian principles to calibrate their judgments by using distributional information about stimulus values within a stimulus series to establish the prior probability of a value because the prior probability is that a member of the smaller series will be smaller than a member of the larger series.

The significance of these studies is magnified by the fact that previous work has not directly demonstrated regression toward a category prototype under conditions in which exemplars from multiple categories were presented to participants. Although Huttenlocher et al. (1991) demonstrated landmark effects in the location of dots in a circle that appeared to be produced by the imposition of vertical and horizontal axes of reference, in previous tests of the model participants were not presented with alternative categories (Huttenlocher, Hedges, Engebretson, & Vevea, 1995; Huttenlocher et al., 2000). Instead, participants have been presented with "a set of stimuli that varied along a single dimension, forming one cluster over a range of values" (Huttenlocher et al., 2000, p. 225) and have been asked to reproduce individual stimuli in each of the experiments (Huttenlocher et al., 2000). The problem with these tests is that they demonstrate the effect of recently presented stimuli on memory for a stimulus but do not directly demonstrate the use of categories. If categories are useful in reconstructing a stimulus, the probability of specific values must differ as a contingency of category membership from the probability of those values in general. This claim can be directly tested only when alternative categories are readily available and the probability of specific values varies as a function of category.

The goal of the present study is to provide such a test of the category adjustment model. In each of the following experiments, participants were briefly exposed to one of a handful of stimuli and asked to reproduce its size on each trial. A simple categorical distinction based on color or shape was used to make specific values more or less likely, given each category. If participants use categories to calibrate their judgments, their reproductions should differ for stimuli with identical values but different category membership. In contrast, if the contextual influence of other stimuli is primarily a function of temporal contiguity, then reproductions should not systematically vary as a function of category membership.

EXPERIMENT 1

In this experiment, we tested the claim that people use category prototypes to calibrate their judgments by presenting participants with a set of blue squares and a set of red squares whose members varied in size over two different but overlapping size ranges. To be more specific, the three largest blue squares were the same size as the three smallest red squares. According to the category adjustment model, classification of the squares into red and blue categories should lead participants to underestimate the size of these blue squares as they adjust them down toward the mean of the smaller blue category and overestimate the size of the same-sized red squares as they adjust them up toward the mean of the larger red squares.

Method

Participants. Thirty-one undergraduates at Lehman College participated for partial course credit in an introductory psychology class.

Materials. The stimuli were square figures that were presented on a 14- or 15-in. VGA monitor set to a 640 × 480 pixel resolution. Pixels were the units used to define stimulus values.

The stimuli consisted of a set of six smaller blue squares whose widths were 10, 30, 50, 90, 100, and 110 pixels and a set of six larger red squares whose widths were 90, 100, 110, 150, 170, and 190 pixels.

Procedure. The procedure was designed to closely match the original procedure of Huttenlocher et al. (2000). Each trial began with a fixation sign (e.g., "+") that appeared on the left side of the monitor and was followed by a square that appeared for 2 sec. After a 1-sec delay, a 5-pixel response square of the same color as the target square appeared in the middle of the screen, and the participants adjusted its size using the two keys on the mouse until they were satisfied that it matched the size of the original target square. We used a response square whose color matched the color of the target square because we wanted to isolate any effect of color category on the initial valuation of the target or the perceived magnitude of the response square from changes in the remembered size of a target that could be attributed to the reconstructive processes postulated by the category adjustment model. Responses were measured in pixels.

The participants were presented with six blocks of experimental trials and a single block of practice trials using the same materials. Each stimulus was presented once in a block, and the order of presentation was randomized within each block. At the end of each block, the participants were informed of the average error of their estimates as measured by the percent deviation of each response from the target square size.

On the last two trials, the participants were instructed to estimate the average size of the red squares and the average size of the blue squares to determine whether each participant had identified the size difference between the two sets of squares. They were told that if they saw a red response square they were to adjust it until it matched their estimate of the average size of the red squares, and if they saw a blue response square they were to adjust it until it matched the average size of the blue squares that they had seen in the experiment. The order of these two trials was randomized.

Results

Prior to calculating a mean for any of the 12 squares, the smallest 1% and the largest 1% of responses in the data set were eliminated for each of the 12 different squares. The means and confidence intervals are presented in Table 1.

Table 1
Mean Estimated Sizes and Confidence Intervals for the 12 Squares in Experiment 1

Size (pixels)	Estimated Size	Confidence Interval	
		Lower Boundary*	Upper Boundary
	Blue Squares		
10	14.89	13.27	16.51
30	37.66	34.99	40.33
50	56.55	53.57	59.52
90	89.02	86.11	91.93
100	96.65	93.61	99.69
110	103.81	100.38	107.24
	Red Squares		
90	89.44	86.59	92.29
100	96.61	93.49	99.73
110	104.78	101.27	108.29
150	136.28	131.56	141.00
170	155.23	150.06	160.40
190	177.77	172.54	183.00

*For a two-tailed test with 27 *df*, $p < .05$.

Availability of prior probabilities. According to the category adjustment model, participants should underestimate the size of the three largest blue squares but overestimate the size of the three smallest red squares. An important element of this prediction is that the participants understood the differences in the distribution of sizes found in the two categories. A sign test conducted on the average estimates provided by the participants at the end of the experiment indicates that the 28 participants who estimated that the red squares were larger than the blue squares constituted a reliable majority ($p < .001$). In addition, the geometric mean of the estimated average size of the red squares (125 pixels) was substantially larger than the geometric mean of that of the blue squares [78 pixels; $F(1,30) = 34.7$, $p < .001$]. Although this estimate for the blue squares is quite a bit greater than the geometric mean of this series (50 pixels) the estimate for the red squares is quite close to the geometric mean of the red squares (130 pixels).[2] These results suggest that the participants associated very different prior probabilities with the three critical blue and red test squares.

Bias. To test the hypothesis that remembered sizes are partially constructed from the category means, a 2 (category: red vs. blue) × 3 (size) repeated measures analysis of variance (ANOVA) was conducted on means of the estimates that the participants made for the 90-, 100-, and 110-pixel squares in each of the two color categories. We restricted this analysis to the 28 participants whose estimates of the average size of the red squares was larger than their estimates of the average size of the blue squares to ensure that the analysis was based on the responses of participants who knew that the blue squares were generally smaller than the red squares. Contrary to the predictions of the category adjustment model, an inspection of Table 1 reveals extremely similar estimates for the blue and red squares. In fact, the mean estimate for the red squares (97 pixels) was just 1 pixel larger than that for the blue squares (96 pixels; $F < 1$). The apparent similarity of these two means is bolstered by the fact that the within-subjects confidence interval for these means is ±1.42 pixels (Loftus & Masson, 1994). Thus, the population mean for the red squares is quite unlikely to differ by more than several pixels from the population mean of the blue squares. The only reliable difference in the means was an increase in estimates with increasing size of the target square [$F(2,54) = 123.0$, $p < .001$].

Although estimates of the critical blue and red test squares did not appear to differ, there is evidence of systematic biases for the overall set of stimuli. An inspection of Table 1 reveals a substantial bias for estimates of the three smallest (i.e., 10, 30, and 50 pixels) and the three largest (i.e., 150, 170, and 190 pixels) squares. The actual size of the three smallest squares is below the lower boundary of the confidence interval for the estimated size of these squares. Similarly, the actual size of the three largest squares is substantially above the upper boundary of the confidence intervals of the estimated

size of each of these squares. This pattern suggests that the participants regressed over the whole range of square sizes covered in the two color categories. The geometric mean of the combined series is 80 pixels. If the participants regressed to this overall mean, then the smallest squares would be overestimated, the middle three sizes just slightly underestimated, and the three largest squares markedly underestimated. This description closely matches the results of the experiment.

If correct, the conclusion that the participants regressed to the overall mean is problematic for the category adjustment model. Although the category adjustment model predicts regression to the overall mean if the participants used the single category "square" for all of the stimuli, it is difficult to see how they learned that red squares were generally larger than blue squares if they were not categorizing squares by color. Moreover, the only evidence to suggest that the participants may have placed red and blue squares in the same category of "square" is the pattern of bias, and it is exactly this pattern that the theory attempts to explain in the first place. Therefore, it seems more plausible that the participants failed to use the differing distributions of red and blue squares not because they categorized these squares equivalently but because they regressed to the mean of a context that was broader than either of the two categories.

EXPERIMENT 2

In Experiment 1, it appeared that estimates regressed toward the mean of the combined distributions rather than toward the means of each of the two categories. Experiment 2 was designed to provide a stronger test of the category adjustment model in two respects. First, the inclusion of only one square that was common to both categories reduced the amount of overlap between the values of the two categories. This change makes the use of category-level information more reliable because the variability of values within each category is reduced relative to the variability of values in the combined stimulus distribution. Although the category adjustment model does not claim that the use of category-level information is strategic, it seems reasonable to assume that this change would make it more difficult for participants to ignore the category membership of squares.

Second, the present design makes it possible to contrast the effect of the combined stimulus range with the effect of the range of values within each category. In Experiment 1, the 90-, 100-, and 110-pixel red squares were smaller than their category mean but slightly larger than the geometric mean of the combined stimulus distribution, which was 80 pixels. The stimulus set was redesigned for Experiment 2 so that the 90-, 100-, and 110-pixel blue squares were smaller than the geometric mean of the combined stimulus distribution but larger than their category mean. If judgments are influenced by the mean of the appropriate color category, then the estimates of these blue squares in this experiment should be smaller than those of the same-sized red squares in Experiment 1 because they should regress down, rather than up, toward their category mean. However, if estimates are influenced primarily by the mean of the combined stimulus distribution, then the estimates for these blue squares should be larger than the estimates for the same-sized red squares because the blue squares should regress up, rather than down, toward the mean of the combined stimulus distribution.

Method

Participants. There were 31 participants from the same population as those in Experiment 1.

Materials. The stimuli consisted of a set of six smaller blue squares, whose widths were 30, 50, 70, 90, 100, and 110 pixels, and a set of six larger red squares, whose widths were 110, 150, 200, 210, 230, and 250 pixels. The geometric mean of the combined stimulus distribution was 112 pixels.

Procedure. The procedure was identical to the procedure in Experiment 1.

Results

The use of the same procedure as in Experiment 1 eliminated outliers. The means and confidence intervals are presented in Table 2.

Availability of prior probabilities. An initial analysis was performed on the estimates of the average sizes of the red and blue squares that the participants made at the end of the experiment. Twenty-five of the participants provided a larger estimate for the red squares than for the blue squares ($p < .001$). The geometric mean of the estimated average size of the red squares (126 pixels) was substantially larger than the geometric means of that of the blue squares [76 pixels; $F(1,30) = 26.0, p < .001$]. This large difference in the estimates suggests that the participants associated very different prior probabilities with the 110-pixel blue and red squares. Subsequent analyses were restricted to the data of the 25 participants who provided a larger estimate for the average size of the red squares.

Table 2
Mean Estimated Sizes and Confidence Intervals for the 12 Squares in Experiment 2

Size (pixels)	Estimated Size	Confidence Interval	
		Lower Boundary*	Upper Boundary
	Blue Squares		
30	42.49	38.74	46.24
50	61.43	58.40	64.46
70	76.22	73.27	79.17
90	93.68	90.26	97.10
100	101.61	96.99	106.22
110	109.15	104.68	113.62
	Red Squares		
110	107.96	103.39	112.53
150	139.11	132.99	145.23
200	178.36	170.14	186.58
210	187.80	179.23	196.37
230	205.62	196.31	214.93
250	222.76	211.31	234.21

*For a two-tailed test with 24 *df*, $p < .05$.

Bias. If the participants failed to use category-level information in Experiment 1 because the overlap between the two categories was too great, then it ought to be possible to observe the use of this information in the present experiment. To test this prediction, we compared the estimate of the size of the blue 110-pixel square with that of the red one. In consistency with Experiment 1, there does not appear to be a difference between the two estimates. The mean of the estimate of the blue square was 109 pixels, and the mean of that of the red square was 108 pixels ($F < 1$). Thus, the degree of overlap in the two categories does not appear to determine whether or not the participants' estimates will regress to the mean of each category.

A second prediction of the category adjustment model is that participants should regress to the mean of each category rather than to the mean of the combined stimulus distribution if they are classifying red and blue squares into different categories. According to this claim, the estimated sizes of the 90-, 100-, and 110-pixel squares should be underestimated relative to the estimates of the comparable red squares provided by the participants in Experiment 1. To compare the participants' estimates of the sizes of these three red squares in Experiment 1 with those of the same-sized blue squares that were provided by the participants in the present experiment, a repeated measures ANOVA with condition as a between-subjects factor and size as a within-subjects factor was used. This analysis revealed that the participants in this experiment provided larger estimates ($M = 101$ pixels) than did the participants in Experiment 1 ($M = 97$ pixels) even though these blue squares were the largest members of their category in the present experiment [$F(1,51) = 4.0, p = .05$]. This result indicates that the participants regressed to the mean of the combined stimulus distribution rather than to the mean of each color category. Finally, the estimates provided for the 90-, 100-, and 110-pixel squares (Ms = 91, 99, and 107 pixels, respectively) differed [$F(2,102) = 122.5, p < .001$]. The interaction between size and condition was not significant ($F < 1$).

The conclusion that the participants regressed to the mean of the combined distribution is also supported by an examination of the mean responses for each square size that are presented in Table 2. This examination indicates that the participants overestimated squares that were smaller than 110 pixels and underestimated squares that were larger than 100 pixels. This pattern is quite consistent with the pattern of regression that was observed in Experiment 1, and it exactly matches what one would expect if the participants regressed to a mean that was slightly smaller than the geometric mean of the combined stimulus distribution ($M = 112$ pixels).

Sequential effects. The degree to which estimates are biased is a function of the standard deviation of the prior distribution or variability of values within a category according to the category adjustment model. This view implies that bias should be a function of the long-run probability of a given category value rather than a function of a response to immediate experience. For example, sampling several large values should not change one's assessment of the probability of sampling smaller values in the immediate future for a reasonably well-established category. Contrary to this view, a rather large body of literature has demonstrated that judgments of a current stimulus frequently assimilate to judgments of an immediately preceding stimulus (DeCarlo, 1992; King & Lockhead, 1981; Staddon, King, & Lockhead, 1980). Post hoc analyses of the data from Experiments 1 and 2 suggest that the participants were fairly sensitive to the magnitude of stimuli encountered in the immediate past.

These analyses were performed on the one square size common to both categories in Experiment 2 and the three square sizes common to both categories in Experiment 1. Estimates were divided into (1) those preceded by a trial in which the square was among the smaller half of the squares in the series and (2) those preceded by a trial in which the square was among the larger half of the squares in the series. In Experiment 1, estimates for the sizes of squares preceded by a small square ($M = 93.8$ pixels) were reliably smaller than estimates for squares preceded by a large square [$M = 99.3$ pixels; $F(1,27) = 32.1, p < .001$]. The interaction between target square size and the size of the square on the preceding trial was not significant ($F < 1$). The data from Experiment 2 also revealed that estimates for squares preceded by a small square ($M = 104.7$ pixels) were smaller than estimates for those preceded by a large square [$M = 112.3$ pixels; $F(1,24) = 16.8, p < .001$]. Although these analyses are post hoc, they strongly suggest that the participants were regressing to the local context rather than to long-run probabilities established by a color category or the combined distribution.[3]

EXPERIMENT 3

In each of the two preceding experiments, participants provided estimates of the average size of squares that differed across two color categories. Although these average estimates suggest that the participants developed very different knowledge about the sizes of blue and red squares, they do not demonstrate that they possessed detailed knowledge about the relative frequencies of different sizes in each category. The goal of the present experiment is to determine the degree to which the participants generated detailed knowledge about the relative frequency of sizes in each color category.

In Experiment 3, we presented participants with squares of six different sizes. For one color (i.e., the negatively skewed color), the frequency of presentation of each size increased with increasing size, but for the second color (i.e., the positively skewed color) frequency decreased with size. Although frequency varied as a function of size for each color, the combination of the two color categories formed a uniform distribution in which frequency of occurrence was constant for each square size.

Previous research has demonstrated that the judged frequency or familiarity of a stimulus is a function of both the frequency with which the target stimulus has been presented and the similarity of the stimulus to other,

previously encountered stimuli (Jones & Heit, 1993; Nosofsky, 1991). These findings suggest that the familiarity of stimuli encountered in the size reproduction task should depend on the degree to which color was encoded with size information. In a final transfer task, we asked participants to judge which member of each pair of squares in a series had appeared more frequently during these size reproductions. If participants encode both the size and the color of a square, then they should prefer the square whose frequency is greatest given its color. In general, this influence of size and color should produce a preference for larger squares of the negatively skewed color and a preference for smaller squares of the positively skewed color. In contrast, if participants encode size but not color for a stimulus, they should have no preference for either of two square sizes because all sizes appeared with equal frequency.

Table 3
Mean Estimated Sizes and Confidence Intervals for the 10 Squares in Experiment 3

Size (pixels)	Estimated Size	Confidence Interval	
		Lower Boundary*	Upper Boundary
	Small Squares		
30	38.33	37.16	39.50
50	58.30	56.80	59.80
70	75.41	73.62	77.19
100	101.93	99.61	104.25
120	117.16	113.54	121.65
	Large Squares		
50	57.96	55.87	60.05
70	78.06	75.96	80.16
100	102.26	99.76	104.76
120	118.54	116.01	121.07
150	150.55	147.87	153.24

*For a two-tailed test with 51 *df*, $p < .05$.

Method

Participants. Fifty-two participants were drawn from the same population as those in Experiment 1. The color assigned to the positively skewed distribution was red for 25 participants and blue for 27 participants.

Materials. The participants were presented with squares of 30-, 50-, 70-, 100-, 120-, and 150-pixel widths. Half of these squares were blue, and half were red. Across both color categories, each size appeared with equal frequency, but within each color category the relative frequency was positively skewed for one color and negatively skewed for the other. Thus, in a given block the 30-, 50-, 70-, 100-, 120-, and 150-pixel squares appeared with a frequency of 4, 3, 2, 2, 1, and 0 times, respectively, in one color but with a frequency of 0, 1, 2, 2, 3, and 4 times, respectively, in the other color. The positively skewed color will be referred to as the *small category*, and the negatively skewed color will be referred to as the *large category*.

Fifteen pairs of squares were constructed for the transfer test. Each of eight pairs was constructed by pairing a red and a blue square of the same size. These pairs included pairs formed from squares of the six sizes presented in the size judgment task and from squares of two additional sizes that had not been previously presented (40 and 130 pixels). In addition to these same-sized pairs, seven pairs of same-color squares whose members differed in size were included.

Procedure. The procedure differed from that of the previous experiments in three respects. First, the participants were explicitly instructed that they would view figures drawn from two different categories: blue squares and red squares. Second, the participants completed two blocks of size reproduction trials. Within each block, the order of stimuli was randomized. Third, instead of producing the category average for each set of squares, the participants judged which member of a pair of squares was "more familiar because it had appeared more frequently" during the size judgments. The materials for this task consisted of one block of 15 pairwise judgments. The order of pairs and the ordering of the two members of each pair were randomized over participants, with the restriction that half of the judgments were of pairs whose more frequently seen member appeared on the left.

Results

Outliers were eliminated through use of the same procedure that was used in previous experiments. The means and confidence intervals are presented in Table 3.

Availability of prior probabilities. The main focus of this experiment was to determine whether the participants encoded both size and color information for the stimuli as they remembered and reproduced the size of stimuli presented in the first part of the experiment. Two measures were calculated from the participants' responses on the frequency comparison task. The first measure was simply the proportion of trials on which they chose the more frequent member of the pair. The calculation of this measure did not include the two pairs whose members had equal frequencies (e.g., the pair consisting of one 70-pixel red and one 70-pixel blue square, and the pair consisting of one 100-pixel red and one 100-pixel blue square), but it did include the two pairs of novel sizes (e.g., the pair consisting of one 40-pixel red and one 40-pixel blue square, and the pair consisting of one 130-pixel red and one 130-pixel blue square). A choice of the positively skewed color for the 40-pixel squares and a choice of the negatively skewed color for the 130-pixel squares were counted as correct. The mean proportion correct was .64, which was shown to be highly significant by a sign test ($p < .001$). In addition, 41 of the 52 participants had accuracies greater than .50 ($p < .001$).

The second measure was designed to assess sensitivity to the relative frequency with which each of member of a pair was presented. For each pair, the ratio of the frequency of the more common item to the combined frequency of both items was calculated. This measure did not include the two pairs with novel sizes whose frequency of presentation was zero, but it did include the two pairs with equal-frequency items. The correlation between this measure and the average proportion correct for each pair was $r(13) = .67$ ($p < .05$).

Finally, an analysis of the two novel stimuli was conducted in which the frequency of choices of the small category for the 40-pixel square and the large category for the 130-pixel square was tabulated. The participants made these choices on 63% of the 104 trials in which one of these two pairs appeared, which was found to be significant by a sign test ($p < .01$). Although the previous re-

sults could be attributed to experience with the actual stimuli, the latter result with novel sizes strongly suggests that the participants were influenced by the similarity of a stimulus to other stimuli in terms of both color and size.

Bias. As in the previous two experiments, the reproduced sizes of squares from the smaller category and of those from the larger category were compared using a repeated measures ANOVA in which the size of the stimulus (i.e., 50, 70, 100, or 120 pixels) and its category (i.e., small or large) were within-subjects factors. As was mentioned previously, the assignment of red and blue to the small and large categories was counterbalanced in this experiment. An initial ANOVA in which this assignment of color to a category was included along with the size of the color category (i.e., large or small) and stimulus size as factors did not reveal any effect of the specific color that was assigned to the categories, so the data were collapsed over the two assignments of color to category (e.g., blue–small and red–large vs. red–small and blue–large).

As in the first two experiments, the assignment of stimuli to a category did not produce any reliable differences. Reproductions of the stimuli from the smaller category were just slightly smaller ($M = 88$) than those of the stimuli from the larger category [$M = 89$; $F(1,51) = 1.6$, $p > .05$]. The size of the stimulus produced reliable differences in responses [$F(3,153) = 1{,}050.8$, $p < .001$]. Finally, category did not interact with stimulus size [$F(3,153) = 1.0$, $p > .05$].

Although category membership did not appear to influence size reproductions, the reproductions do exhibit systematic biases. Confidence intervals were established by calculating the mean reproduction for each size for each participant and calculating the standard error of the mean for each size. The lower boundaries of the confidence intervals for each of the smallest three squares (i.e., 30, 50, and 70 pixels) were above the actual sizes of the squares, indicating that responses to these three stimuli were reliably biased toward the mean of the combined distribution. In contrast, the actual size of each square fell within the confidence interval of each of the three largest squares. In fact, only the response to the 120-pixel square was smaller than the actual value of the square. Thus, the pattern of bias in this experiment did not completely replicate the pattern of bias that was observed in the first two experiments.

Sequential effects. In Experiments 1 and 2, the magnitude of a response was affected by the size of the square on the previous trial. A similar analysis was conducted on these data by comparing the reproduced sizes of squares when they were preceded by one of the three smallest squares with the sizes of the same squares when they were preceded by one of the three largest squares. This analysis was restricted to reproductions of the four middle-sized squares, which were common to both categories. In consistency with the results from the two previous experiments, responses to targets preceded by a large square ($M = 91.6$) were greater than responses to targets preceded by a small square [$M = 85.9$; $F(1,51) = 51.7$, $MS_e = 63.9$, $p < .001$]. This effect of the size of the preceding square did not interact with the effect of target size [$F(3,153) = 1.7$, $MS_e = 47.0$, $p = .17$].

EXPERIMENT 4

A common procedure in the literature on concept formation and classification with artificial categories is to present participants with examples from two different categories and ask them to classify each item into one of them. A variety of results, such as sensitivity to the relative frequency of attributes within a category (Medin & Schaffer, 1978; Rosch & Mervis, 1975), suggest that participants are able to induce categories through this procedure. Although the results of several manipulation checks in the preceding experiments indicate that the participants were sensitive to the correlation between size and color, these experiments do not provide direct evidence that the participants categorized the stimuli into separate categories for red and for blue squares. The goal of the present experiment was to require participants to learn this categorical distinction in an initial category-learning phase to determine whether the effects of category membership would be apparent in memory for size if the category had been previously established in this manner.

Method

Participants. Forty-one participants were drawn from the same population as in Experiment 1. Two participants were dropped from the analyses because they did not complete the second task.

Materials. The stimuli for both tasks consisted of a set of six smaller blue squares whose widths were 30, 50, 70, 90, 100, and 110 pixels and a set of six larger red squares whose widths were 100, 110, 150, 200, 210, and 230 pixels.

Procedure. In the initial category-learning task, the participants were instructed that they would see figures that belonged to one of two categories and that they were to try to learn the category membership of these figures by paying attention to the feedback provided about their choices. On each trial, a figure appeared on the screen until the participant pressed one of two keys on the keyboard: the key labeled "A" or the key labeled "B." RT was measured from the onset of the square. If the participant's choice was correct, the word READY appeared on the screen to indicate the start of the next trial. If the participant's choice was incorrect, the words WRONG ANSWER appeared on the screen for 2.5 sec prior to the start of the next trial. The participants received an initial block of practice trials, followed by six additional blocks of experimental trials. The order of stimuli was randomized within each block.

At the outset of the size reproduction task, the participants were instructed that they would be presented with the same kind of items as in the previous task but that in this phase of the experiment they would be asked to remember and reproduce from memory each of these figures. In all other respects, the procedure for the size reproduction task was identical to that of Experiment 1.

Results

Categorization. In addition to determining that the participants were able to learn the color categories, we analyzed the categorization data to determine whether the size and/or color of the square influenced performance. An initial inspection of the data indicated that overall accuracy was high ($M = .97$) and fairly stable after the third block of trials. For this reason, the analy-

ses of the categorization data were conducted on the last four blocks of trials. If the participants learned the association of size and each of the two color categories, then responses to the 100- and 110-pixel squares should be slower and less accurate than responses to smaller or larger squares because size is not a cue to category membership. On the other hand, the smaller or larger squares should be slightly easier to categorize because size as well as color should provide a basis for correctly categorizing these more extreme squares. To test this prediction, responses to the red and blue 100- and 110-pixel squares were aggregated to determine the mean latency and accuracy on these trials, and the data for the remaining smaller and larger squares were aggregated to determine their mean accuracy and latency. Mean accuracy and latency were calculated for these two groups of trials for each participant for each block.

In separate analyses, the accuracy and latency data for these two groups of squares were compared using a repeated measures ANOVA in which the two size groups (i.e., moderate vs. extreme) and block were within-subjects factors. Although the RTs were slightly greater for the moderate squares ($M = 775$) than for the extreme squares ($M = 727$), this difference was not reliable [$F(1,38) = 2.2, p = .15$]. However, performance on moderate squares ($M = .96$) was reliably less accurate than performance on extreme squares [$M = .99$; $F(1,38) = 4.93, p < .05$]. This influence of square size did not systematically differ across blocks for RT ($F < 1$) or accuracy [$F(3,114) = 2.17, p = .1$]. The only other reliable effect in the analysis of the RT or accuracy data was a decline in RT over blocks [$F(3,114) = 6.17, p < .01$]. Taken together, these results suggest that the participants were slightly more accurate at classifying squares as belonging to the red or the blue category when size was a valid cue to category membership, and this difference could not be attributed to a speed–accuracy trade-off. The fact that this advantage for extreme squares occurred even when accuracy was relatively high suggests that this advantage persisted even after the participants had acquired the categories.

Bias. As in the previous experiments, removal of the smallest and largest 1% of the responses for each stimulus eliminated outliers. The means and confidence intervals are presented in Table 4.

The reproduced sizes of squares from the smaller and larger categories were compared using a repeated measures ANOVA in which the size of the stimulus (e.g., 100 vs. 110 pixels) and its category (e.g., small vs. large) were within-subjects factors. Once again, the assignment of stimuli to a category did not produce any reliable differences. The difference between the reproductions for the smaller category ($M = 104.2$) and those for the larger category ($M = 104.6$) was less than 1 pixel ($F < 1$). Stimulus size produced reliable differences in responses [$F(1,38) = 97.1, p < .001$]. Finally, category did not interact with stimulus size ($F < 1$).

Although category membership did not appear to influence size reproductions, the reproductions do exhibit systematic biases. An inspection of Table 4 reveals that the lower boundaries of the confidence intervals for each of the smallest three squares (e.g., 30, 50, and 70 pixels) were greater than the actual size of the squares, indicating that responses to these three stimuli were reliably biased toward the mean of the combined distribution. Similarly, the actual sizes of the four largest squares were above the upper boundaries of their respective confidence intervals, indicating that they were reliably biased toward the mean of the combined distribution. Thus, the pattern of bias in this experiment was quite similar to those observed in Experiments 1 and 2.

Table 4
Mean Estimated Sizes and Confidence Intervals for the 12 Squares in Experiment 4

Size (pixels)	Estimated Size	Confidence Interval	
		Lower Boundary*	Upper Boundary
	Blue Squares		
30	39.33	37.23	41.43
50	58.01	55.90	60.12
70	75.46	72.83	78.10
90	93.03	90.05	96.00
100	100.36	97.21	103.51
110	108.12	104.95	111.29
	Red Squares		
100	100.14	96.90	103.37
110	109.00	105.64	112.33
150	140.00	136.32	143.69
200	183.44	178.53	188.35
210	192.91	187.80	198.03
230	211.50	205.85	217.15

*For a two-tailed test with 38 *df*, $p < .05$.

Sequential effects. Responses to the two sizes that were common to both categories (i.e., 100 and 110 pixels) were analyzed to determine whether or not they were influenced by the size of the preceding square. The mean for trials preceded by one of the four smallest squares was compared with the mean for trials preceded by one of the four largest squares. In consistency with the results of the previous experiments, responses to targets preceded by a large square ($M = 107.3$) were greater than responses to targets preceded by a small square [$M = 101.6$; $F(1,38) = 45.9$, $MS_e = 13.7$, $p < .001$].

EXPERIMENT 5

In the previous experiments, estimates generally regressed toward the mean of the combined distribution rather than toward the mean of each of the two color categories. This finding is inconsistent with the category adjustment model unless the participants ignored the color differences between the two sets of squares and induced a single category.

In an attempt to provide an even stronger manipulation of category membership, the participants were asked to reproduce figures that were either blue circles or red

squares. According to Rosch, Mervis, Gray, Johnson, and Boyes-Braem (1976), categories are hierarchically organized and exemplars can be classified at different levels of abstraction within this hierarchical organization. Although items can be classified at several levels, Rosch et al. proposed that they are most readily classified at the basic level (e.g., *guitar*, *piano*) rather than at the superordinate (e.g., *musical instrument*) or subordinate (e.g., *folk guitar*, *grand piano*) level. For example, adults are much more likely to use basic-level descriptors than super- or subordinate-level descriptors to name pictures. This analysis suggests that circles and squares can be classified at the superordinate level as geometric shapes, at the basic level as circles and squares, and finally at the subordinate level as red squares or blue squares. More important, it suggests that the most salient categorization should be between squares and circles.

In this experiment, the participants in the multiple-category condition reproduced figures taken from a set of small blue circles and a larger set of red squares, whereas the participants in the single-category condition reproduced figures taken only from the set of small blue circles. The participants in the multiple-category condition were informed that they would see shapes that belonged to one of two categories: blue circles and red squares. If the participants in the multiple-category condition use separate categories for circles and squares to reconstruct exemplars drawn from each of these two categories, then their responses should be biased with respect to the characteristics of the category from which the exemplar was drawn rather than with respect to the characteristics of members drawn from other categories. Thus, their responses to circles should be extremely similar to the responses of the participants in the single-category condition. In contrast, if memory for a stimulus is influenced by the presence of stimuli from other categories, then the presence of larger squares should lead to the production of larger responses to the circles in the multiple-category condition than in the single-category condition.

Method

Participants. Fifty-seven participants in the single-category condition and 53 participants in the multiple-category condition were drawn from the same population as in Experiment 1. One participant was dropped from the single-category condition for producing 15 outliers (31%).

Materials. In both conditions, the participants were presented with a set of four blue circles whose diameters were 20, 30, 50, and 90 pixels. The second category in the multiple-category condition consisted of four larger red squares with widths of 110, 150, 170, and 190 pixels.

Procedure. The procedure was identical to that of Experiment 1 with three exceptions. First, the participants in the multiple-category condition were explicitly informed that they would view figures drawn from two different categories: blue circles and red squares. Second, the response figure was changed to a 5-pixel-diameter circle when the target figure was a circle in the multiple-category condition. Third, the participants completed 7 blocks of experimental trials in the multiple-category condition and 13 blocks of experimental trials in the single-category condition.

Results

Outliers were eliminated through use of the same procedure that was used in the previous experiments. The means and standard deviations are presented in Table 5.

Bias. According to the category adjustment model, category-level information is used to guide the reproduction of a stimulus from memory and responses should be biased toward the mean of the category to which they belong. This prediction was tested by conducting a repeated measures ANOVA on the means of the circle responses with condition as a between-subjects factor and size as a within-subjects factor. Contrary to the predictions of the category adjustment model, responses to circles in the multiple-category condition ($M = 51.1$) were larger than those in the single-category condition [$M = 48.1$; $F(1,107) = 23.9$, $MS_e = 43.2$, $p < .01$]. In addition, the responses provided for the four different sizes (i.e., 20-, 30-, 50-, and 90-pixel circles) differed [$F(3,321) = 4,926$, $p < .001$]. Although the mean responses in the multiple-category condition were larger for all four sizes, the interaction between size and condition was significant [$F(3,321) = 4.44$, $MS_e = 15.16$, $p < .01$].

Sequential effects. In the previous experiments, the magnitude of a response was affected by the size of the square on the previous trial. In the multiple-category condition of the present experiment, a figure could be preceded by a circle or by a square. If the size of the preceding stimulus had an effect on responses to the target stimulus, then responses for each of the circles should be larger when the circles were preceded by a square than when they were preceded by a circle, because each of the squares was larger than any of the circles. Consistent with the results of the two previous experiments, responses to targets preceded by a square ($M = 52.6$) were greater than responses to targets preceded by a circle [$M = 49.6$; $F(1,46) = 39.0$, $MS_e = 21.65$, $p < .001$].[4] Although responses were greater following squares for all of the targets, there was an interaction between the size of the

Table 5
Mean Estimated Sizes and Mean Standard Deviations (*SD*s) for the Figures in Experiment 5

Size (pixels)	Estimated Size	*SD*
Single Category Condition		
20	24.37	2.78
30	33.48	3.53
50	49.55	5.06
90	84.88	7.96
Multiple Category Condition		
20	27.19	3.88
30	37.24	4.51
50	54.27	5.42
90	85.91	8.00
110	104.80	11.40
150	135.70	14.80
170	154.60	16.70
190	178.80	16.70

stimulus on the preceding trial and the influence of the prior stimulus on the current response [$F(3,138) = 12.6$, $MS_e = 16.35$, $p < .01$].

Variability of responses. In addition to predicting that bias should occur when category information is used to calibrate memory for an individual stimulus, the category adjustment model claims that the variability of responses to stimuli should be a function of the uncertainty about the true value of the stimulus and the uncertainty of the category prototype as an estimate of the prior probability of the stimulus. More specifically, the standard deviation (S) of responses (R) is

$$S(R) = \lambda\sigma_M,$$

where λ is the weight given to the memory of a stimulus and σ_M represents the inexactness of the representation of the stimulus in memory. It is important to point out that λ is a smooth monotonic decreasing function of the ratio of memory uncertainty (σ_M) to the inexactness of the central value of the category (σ_C) with a range of 0 to 1. Thus, λ and the variability of responses increase as the range of values in a category increases. This predicted relationship between the range of values within a category and the variability of responses to a member of that category is the basis for the claim that the accuracy of responses can be improved by using category-level information when memory for an individual stimulus is inexact. For example, if memory uncertainty for the stimulus is very large but the range of values within the category is generally fairly small, then the uncertainty of memory for the stimulus should be large relative to the uncertainty of the prototype as an a priori predictor of the stimulus value. These circumstances imply that the ratio of σ_M to σ_C would be large and, therefore, the variability of responses, $S(R)$, should be smaller than σ_M because the value of λ would be much less than 1.

Although testing this prediction was not an original goal of the present experiment, a comparison of the variabilities of responses to the circles in the two conditions provides some insight into whether utilizing separate categories in the multiple-category condition could have improved the accuracy of responses in that condition. This comparison was made by calculating the variance of each participant's responses for each of the four circles. These variances were analyzed using a repeated measures ANOVA in which condition was a between-subjects factor and circle size was a within-subjects factor. The results indicate that the variability of responses was smaller in the single-category condition ($M = 4.8$) than in the multiple-category condition [$M = 5.4$; $F(1,107) = 5.3$, $MS_e = 7.91$, $p < .05$]. In addition to this difference between the conditions, variability increased as a function of circle size [$F(3,321) = 136.2$, $MS_e = 3.37$, $p < .001$]. The reduced variability in the single-category condition in comparison with that in the multiple-category condition is important because it indicates that the presence of the squares decreased the accuracy of responses in the multiple-category condition. The fact that the inclusion of the squares reduced accuracy in the multiple-category condition indicates that this manipulation was sufficiently robust to change the participants' behavior if they were guided by Bayesian principles to use category information to calibrate their judgments.

DISCUSSION

The results of these five experiments are inconsistent with the claim that regression to the mean in reproduction tasks is produced by a general use of categorical information to improve the accuracy of memory for individual stimuli. In all of the experiments, the participants appeared to be sensitive to the overall range of stimulus values and to ignore relevant categorical distinctions. Specifically, their responses to squares did not differ as a function of the squares' color even though two independent measures indicated that the participants had knowledge about the correlation between size and category membership. In Experiment 5, the participants' responses to circles in the multiple-category condition were biased by the introduction of items from a different category, and the variability of their responses was greater than that of the responses of participants who were presented with the same set of circles but no squares. Finally, the responses in all of the experiments were significantly affected by the size of the stimulus on the immediately preceding trial.

The category adjustment model is not unique in claiming that regression to the mean is guided by the values of similar stimuli in the current context (Estes, 1997; Hellström, 1985; Helson, 1964). For example, adaptation level (AL) accounts (Helson, 1964; Marks, 1993) claim that the response to a stimulus is a weighted combination of the effects of the stimulus and an AL or value associated with the pooled effect of contextually relevant stimuli. A considerable challenge for accounts of assimilation effects has been to explain which stimuli are "contextually relevant." Thus, one of the appeals of the category adjustment model is that it appears to provide some constraints on which stimuli should be influential by claiming that inductive categories organize recent experience and provide a structure that can be used to calibrate memory for specific episodes. Unfortunately, the present results undermine this appeal because they suggest that the development or use of inductive categories may not be guided by obvious distinctions. For example, Experiment 5 suggests that the presence of stimuli from a different basic-level category will influence reproductions of stimuli from a second category. Without a priori principles for determining which inductive category will be used to calibrate memory for a particular stimulus, the constraint provided by these categories is minimal because one can always construct a category that includes the appropriate stimuli (e.g., geometric shapes presented in the last 45 min) on an ad hoc basis to explain a particular pattern of regression. The present results indicate that the explanatory power of these inductive categories may be as difficult to establish as it was for previous conceptions of context, such as Helson's AL.[5]

A second appeal of the category adjustment model is that it provides a plausible rationale for assimilation toward a central tendency by proposing that this assimilation has the effect of improving the overall accuracy of responses. Several of the findings of this study suggest that the assimilation observed in these experiments was not determined by a strictly Bayesian use of category-level information. First, reliable sequential effects were identified in all four experiments. These effects have the potential to explain much of the behavior predicted by the category adjustment model. As Marks (1993) indicated, sequential effects have the potential to produce net assimilative effects in which judgments of a stimulus are biased toward the contextual set. If stimuli assimilate to the preceding stimulus, a stimulus will be overestimated when it is preceded by a larger stimulus and underestimated when it is preceded by a smaller stimulus. The reason that this sequential influence mimics the predictions of the category adjustment model is that the average impact of these effects depends on the distribution of stimulus values. When a stimulus is the smallest member of a group of stimuli, the effect of the previous stimulus will always be to produce an increase in the estimated value of the stimulus. In contrast, if the same stimulus is the largest, it will always be underestimated because it will always be preceded by a smaller stimulus. The net effect of prior stimuli on a moderately sized stimulus will be reduced because the effects of smaller stimuli will be opposite the effects of larger stimuli. Moreover, this argument can be extended to make a prediction about the effect of the variability of stimulus values on the variability of estimates that is at least qualitatively similar to the predictions of the category adjustment model. When a stimulus is presented in the context of a more variable set of stimulus values, the assimilative influence of a preceding stimulus will be more variable than when the stimulus is presented in a set of less variable stimulus values. Therefore, the presence of sequential effects suggests that one does not need to rely on the use of category-level information to explain patterns of assimilation or response variability.

Although the category adjustment model does not claim that category induction is necessarily guided by Bayesian principles, the influence of squares on the reproduction of circles in Experiment 5 is difficult to reconcile with the general claim that categories are used to calibrate judgment. There is little doubt that the basic-level distinction between squares and circles would figure prominently in behaviors such as naming, sorting, and describing the stimuli used in Experiment 5. If we are to believe the claims of the category adjustment model, the effect of squares on judgments of circles indicates that the participants induced a category in the multiple-category condition that included both blue circles and red squares. This putative use of a combined category is problematic because it implies that the participants ignored categories that had the potential to provide a better source of prior information about a stimulus in favor of a larger and less informative category defined by temporal context. This reliance on temporal contiguity to determine category membership is hard to explain because this kind of ad hoc category should provide a very poor source of prior information about a new stimulus in most circumstances. Thus, the results of Experiment 5 are inconsistent with the general proposition that we use category-level information to the extent that it provides useful base-rate information about individual stimuli.

In summary, there is a great deal of evidence that our memory for the magnitude of an instance is frequently less extreme than the original magnitude of the instance (Huttenlocher et al., 1991; Kerst & Howard, 1978; Moyer et al., 1978). The category adjustment model proposes that this assimilation reflects the use of estimation procedures that maximize the accuracy of estimates through prior probabilities established by category membership. The present results suggest that estimates regress to the mean of a frame of reference that is broader than membership in a specific category. They also suggest that this regression may reflect a conservative response to the immediate context rather than a rational use of the long-run probabilities established by category membership. Further work will be required to determine whether the category adjustment model should be regarded as a truly descriptive model of recollective processes or a normative model to be contrasted with these processes.

REFERENCES

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. London: Cambridge University Press.

DeCarlo, L. T. (1992). Intertrial interval and sequential effects in magnitude scaling. *Journal of Experimental Psychology: Human Perception & Performance*, **18**, 1080-1088.

Estes, W. K. (1997). Processes of memory loss, recovery, and distortion. *Psychological Review,* **104**, 148-169.

Hellström, Å. (1985). The time-order error and its relatives: Mirrors of cognitive processes in comparing. *Psychological Bulletin*, **97**, 35-61.

Helson, H. (1964). *Adaptation-level theory.* New York: Harper & Row.

Hollingworth, H. L. (1910). The central tendency of judgment. *Journal of Philosophy, Psychology, & Scientific Method*, **7**, 461-469.

Huttenlocher, J., & Hedges, L. V. (1992). Reconstructing the past: Category effects in estimation. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 28, pp. 251-280). Orlando, FL: Academic Press.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, **98**, 352-376.

Huttenlocher, J., Hedges, L. [V.], Engebretson, P. H., & Vevea, J. (1995, November). *Reconstructing experiences from memory: The effects of categories.* Paper presented at the 36th Annual Meeting of the Psychonomic Society, Los Angeles.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, **129**, 220-241.

Jones, C. M. M., & Heit, E. (1993). An evaluation of the total similarity principle: Effects of similarity on frequency judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **19**, 799-812.

Kerst, S. M., & Howard, J. H. (1978). Memory psychophysics for visual area and length. *Memory & Cognition*, **6**, 327-335.

King, M. C., & Lockhead, G. R. (1981). Response scales and sequential effects in judgment. *Perception & Psychophysics*, **30**, 599-603.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, **1**, 476-490.
Marks, L. E. (1988). Magnitude estimation and sensory matching. *Perception & Psychophysics,* **43**, 511-525.
Marks, L. E. (1992). The slippery context effect in psychophysics: Intensive, extensive, and qualitative continua. *Perception & Psychophysics*, **51**, 187-198.
Marks, L. E. (1993). Contextual processing of multidimensional and unidimensional auditory stimuli. *Journal of Experimental Psychology: Human Perception & Performance*, **19**, 227-249.
Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.
Moyer, R. S., Bradley, D. R., Sorensen, M. H., Whiting, J. C., & Mansfield, D. P. (1978). Psychophysical functions for perceived and remembered size. *Science*, **200**, 330-332.
Needham, J. G. (1935). The effect of the time interval upon the time-error at different intensive levels. *Journal of Experimental Psychology*, **18**, 539-543.
Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception & Performance*, **17**, 3-27.
Poulton, E. C. (1979). Models for biases in judging sensory magnitude. *Psychological Bulletin,* **86**, 777-803.
Rosch, E., & Mervis, C. B. (1975). Family resemblance studies in the internal structure of categories. *Cognitive Psychology*, **7**, 573-605.
Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, **8**, 382-439.
Schneider, B., & Parker, S. (1990). Does stimulus context affect loudness or only loudness judgment? *Perception & Psychophysics*, **48**, 409-418.
Staddon, J. E. R., King, M., & Lockhead, G. R. (1980). On sequential effects in absolute judgment experiments. *Journal of Experimental Psychology: Human Perception & Performance*, **6**, 290-301.
Wedell, D. H. (1995). Contrast effects in paired comparisons: Evidence for both stimulus-based and response-based processes. *Journal of Experimental Psychology: Human Perception & Performance*, **21**, 1158-1173.
Wedell, D. H. (1996). A constructive-associative model of contextual dependence of unidimensional similarity. *Journal of Experimental Psychology: Human Perception & Performance*, **22**, 634-661.

NOTES

1. In fact, Hellström (1985) makes a similar argument that the estimate of a particular stimulus may be more accurate over the long run if it is weighted by the adaptation level of a stimulus series. The primary difference between his proposal and the category adjustment model is an explicit reliance on Bayes's theorem and category-level information in the category adjustment model.

2. It is difficult to know whether the appropriate comparison for these estimates is the geometric mean or the arithmetic mean of the stimulus series. Wedell (1996) found that the psychophysical function was a linear function of square width for single-stimulus ratings but a negatively accelerated function of square width for pairwise dissimilarity ratings. If the psychophysical function is a negatively accelerated function of square width, then the arithmetic mean will tend to overestimate the average subjective magnitude of the stimulus set relative to the geometric mean, which assumes that the subjective magnitude of each stimulus can be expressed as a function of the log of its physical size.

3. This conclusion is further supported by a reanalysis of the data used to compare the estimated size of the red squares in Experiment 1 with that of the same-sized blue squares in Experiment 2. In this reanalysis, the data were restricted to trials preceded by a stimulus from the set of stimuli common to both experiments (i.e., 30, 50, 90, 100, 110, and 150). The difference between the two conditions was not significant in this reanalysis [$F(1,51) = 2.04$, $p = .16$]. This lack of significance suggests that some of the difference in the original analysis may be due to a difference in local context rather than to a difference in the complete set of stimuli per se.

4. Six participants were dropped for missing data because the random order did not produce an observation for every size that was preceded by both a square and a circle.

5. Although the underlying concepts of an AL and a category prototype differ, the category adjustment model could be viewed as a special case of a more general AL model in which the AL is equivalent to the category prototype and the degree to which the AL is weighted in the reproduction of a stimulus is a monotonically decreasing function of the ratio of the uncertainty of memory and the inexactness of the prototype as an estimate of any category value.

(Manuscript received February 4, 2003;
revision accepted for publication September 18, 2004.)