Journal of Experimental Psychology:
Learning, Memory, and Cognition
1984, Vol. 10, No. 4, 638-648

Evaluation of Exemplar-Based Generalization and the Abstraction of Categorical Information

Jerome R. Busemeyer
Purdue University

Gerald I. Dewey and Douglas L. Medin
University of Illinois at Urbana-Champaign

This article reformulates and reanalyzes a problem originally put forth by Homa, Sterling, and Trepel (1981). The question is whether a pure, exemplar-based abstraction process is an adequate model of category learning or whether it is necessary to postulate an additional prototype-abstraction process. Based on quantitative discrepancies from a pure, exemplar-based model, Homa et al. argued that it was necessary to recognize the operation of a prototype-abstraction process in order to fully explain their results. However, Homa et al. never actually fit the exemplar plus prototype model to the data to determine if indeed the additional prototype process could explain the deviations from the pure exemplar model. The present article compared the pure exemplar model with a mixed (exemplar plus prototype) model and did not find consistent evidence requiring the postulation of an additional prototype-abstraction process. These results point out the difficulty of distinguishing alternative classification models and underscore the need for careful analytic work in this area.

Homa, Sterling, and Trepel (1981) recently reported a pair of experiments using a category-abstraction paradigm that were designed to evaluate exemplar-based models of generalization for ill-defined categories. Subjects learned to classify 35 high-level distortions of prototype patterns into three categories defined by 5, 10, and 20 different patterns and then were given transfer tests either immediately or a week later. The new transfer patterns included old, new, and prototype exemplars. During the transfer test, subjects also were presented a number of unrelated exemplars or foils, and they were instructed to assign these foils to a junk category. The new patterns had one of five levels of objective similarity to a particular, old training stimulus but otherwise were high-level distortions of the category prototype. The primary focus in the study was on the effect of specific new-old similarity as a function of category size and time of the transfer test.

Three different versions of exemplar-based generalization were considered, a single member (SM), a fixed set (FS) model, and a complete set (CS) model. These models differ in the number and nature of the stored exemplars retrieved in response to a test stimulus. The SM model assumes that the stored member most similar to the probe is retrieved from each category, and the probe is placed into the category associated with the best similarity match. The SM model corresponds to a proximity model (Hintzman & Ludlam, 1980; Reed, 1972), and the FS model can be thought of as a generalization where N patterns are retrieved for each category. Homa et al. (1981) assumed that the stored exemplar most similar to the transfer stimulus is always included in the retrieved set of exemplars. Finally, the CS model assumes that all stored patterns are accessed in response to a probe.

To formalize these ideas, Homa et al. (1981) developed prediction equations for each of these models. First let X be the similarity between a test probe and the closest same-category member that is stored, let S_w be the average within-category similarity to the remaining members of the category, and let S_b

This research was supported by U.S. Public Health Service Grant MH 32370.

We wish to thank Don Homa for comments on earlier drafts of this article and for providing tables of the results.

Requests for reprints should be sent to Jerome R. Busemeyer, Psychological Sciences, Purdue University, West Lafayette, Indiana 47907.

be the average between-category similarity of the test probe to members of alternative categories. Then, the classification probability (P) for the CS model, according to Homa et al., would be as follows:

$$P[A|i] = [X + (N_a - 1)S_w]/$$
$$[X + (N_a - 1)S_w + N_bS_b], \quad (1)$$

where $P[A|i]$ is the probability of classifying Pattern i into Category A, N_a is the number of stored patterns in Category A, and N_b is the number of stored patterns in alternative categories.

Then, again following Homa et al., for the FS model Equation 1 reduces to the following:

$$P[A|i] = [X + (N - 1)S_w]/$$
$$[X + (N - 1)S_w + (M - 1)NS_b], \quad (2)$$

where M is the number of categories learned, and N is the number of stored exemplars retrieved per category at the time of classification. Note that $N = 1$ in Equation 2 for the SM model.

From these equations, Homa et al. (1981) stated that the SM and the FS exemplar models could not predict that the effects of old–new similarity would interact with category size. Both studies, however, showed that specific old–new similarity had a smaller effect as category size increased.

They also derived the somewhat counterintuitive result that the CS exemplar model predicts the obtained interaction. Homa et al. (1981) went on to attempt to distinguish the CS model from a mixed model that not only assumes storage of exemplars but also assumes that as category size increases, subjects are more likely to abstract out the central tendency or prototype of a category. They did this by fitting the CS model to the percentage of correct data using Equation 1 and argued that although the CS model provided a reasonable fit (see Table 4 of Homa et al., 1981), there were certain apparent shortcomings. Homa et al. reported that the predicted category size effects were consistently smaller than the observed category size effects for both new and prototype patterns (see Table 5 of Homa et al., 1981). Furthermore, the parameter values for delayed tests seemed to change in a way that was complex if not peculiar. Similarity values for new stimuli closest to an old pattern decreased with delay, whereas values for less similar, new stimuli increased. On the basis of these observations, Homa et al. concluded that the results best support a mixed model involving prototype abstraction and that exemplar models probably only hold for small category sizes and immediate tests.

Purpose of This Article

The Homa et al. (1981) study represents a substantial contribution to the literature on ill-defined categories, if only because it represents one of the few attempts to apply formal models to procedures where stimuli are generated from a prototype and where transfer tests are given immediately and after a delay. In addition, Homa et al. provided suggestive evidence that transfer tests themselves modify category representations. That is, the results on a delayed test were somewhat different when subjects had received an initial test (Experiment 1) than when they had not (Experiment 2). To be sure, test effects have been demonstrated in other subdomains of memory research, but the Homa et al. data made the point that categorization theories need to represent and to describe the effects of tests per se on abstraction.

Our main point of departure from Homa et al. (1981) concerns the strategy that led them to endorse a mixed model. Quantitative predictions are developed for the CS exemplar model, and then discrepancies between predicted and observed data are taken as evidence for the mixed model. The mixed model embodies the CS exemplar model supplemented by a prototype-abstraction process that is assumed to be more likely to form as experience with a category increases. This is a plausible alternative model, but it does not seem in the analytic spirit of the rest of the article not to develop it formally and to evaluate it against alternative possibilities.

Although a model that has the CS model as a special case would doubtless do a reasonable job of fitting the data, it nonetheless may be very informative to set down a specific mixed model and to evaluate its consequences. Some of the implications may be counterintuitive. For example, forming a prototype for the largest category may paradoxically facilitate performance most on the smallest category. This might happen if the large category prototype allowed one to reject small category

patterns as inappropriate for the large category. Homa et al. ended their article with an interesting discussion of overgeneralization errors, and a formalized mixed model could add cogency to this issue.

Any theoretical analysis involves certain simplifying assumptions, and the inadequacies of a model should be evaluated in light of such assumptions. In this article, we argue that certain simplifying assumptions made by Homa et al. (1981) may be responsible for the few shortcomings of the CS model. In particular, subjects in the Homa et al. study were allowed to assign stimuli to a miscellaneous, or junk, category, but Equation 1 for the CS model makes no provision for junk responses (i.e., the sum of the predicted probabilities across the three category sizes 5, 10, and 20 equals one, so the predicted probability of a junk response equals zero). Because subjects made use of the junk category, the CS model necessarily must give a less than perfect account of the data.

The major purpose of this article is to compare a modified CS model, which allows for junk responses, with a more general mixed model, which includes both a prototype model component and a CS model component. It is also important to point out that Homa et al. (1981) based their evaluation of the CS model on the fits to the percentage of correct data alone. This did not require the model to make reasonable predictions concerning the confusions with incorrect categories. In order to provide a more rigorous comparison of the mixed versus CS models, both models were fit to the entire confusion matrix. Finally, we consider the ability of other exemplar models to account for the category size effects and provide some general comments on the problem of contrasting prototype and exemplar models.

Reformulation of CS and Mixed Models

Analysis of Confusion Matrices

In order to apply the CS model to the entire confusion matrix, we employed the biased choice model of Luce (1963) because it has proven to be most effective in past research (Smith, 1980; Townsend & Landon, 1982). The biased choice model contains two sets of parameters: *similarity* and *bias* parameters. The similarity parameters represent the total similarity of a pattern to the members of a given category. The bias parameters represent response tendencies or guessing strategies that are important when unequal base rates or payoffs are employed or when instructions emphasize particular usage of categories.

For the present application, there are four categories: the 5-, 10-, and 20-category sizes, and the junk category. The junk category requires special treatment, which is discussed in the next section, and the present discussion is limited to the remaining three categories. Define $E_{a,i}$ as the total similarity of Pattern i to the members of Category A, where A may be Size 5, 10, or 20. When Pattern i is a member of Category A, then $E_{a,i} = X + (N_a - 1)S_w$; where X is the similarity of Pattern i to the closest member of Category A, S_w is the average within-category similarity, and N_a is the number of members in Category A. When Pattern i is not a member of Category A, then $E_{a,i} = N_a S_b$; where S_b is the average between-category similarity, and N_a is the number of members in Category A. The bias parameter associated with Category A is symbolized as β_a. The probability that Pattern i is placed in Category A (given that A equals either the 5-, 10-, or 20-size category) is as follows:

$$P_{CS}[A|i, A = 5, 10, 20] = \beta_a E_{a,i}/(\beta_5 E_{5,i} + \beta_{10} E_{10,i} + \beta_{20} E_{20,i}). \quad (3)$$

Junk Responding

To incorporate junk responses, one needs some assumption about how they may arise. For example, subjects may assign a pattern to the junk category when no training stimulus is similar enough, when the average similarity is not high enough, or when the average similarity is about equal for all contending categories.

An additional complication is that, in general, one would expect objective similarity to

map onto subjective similarity with some variability. This variability probably would not be independent of whatever process gives rise to junk responses. The problem with incorporating such assumptions into the CS model is that it would add so many parameters to the model that it would be impossible to evaluate.

The alternative approach to junk responses that is presented here is admittedly a considerable oversimplification, but it does have the virtue of adding but a single parameter to the CS model. We assume that subjects assign stimuli to the junk category when the sum of the similarities of training stimuli to the probe stimulus is not sufficiently large. One may imagine some similarity threshold that must be exceeded before a probe is classified into one of the three training categories. One can approximate the effects of this hypothetical threshold by an exponential function. Specifically, Equation 3 is modified as follows:

$$P_{CS}[A|i] = P_{CS}[A|i, A = 5, 10, 20][1 - e^{-\alpha T}], \quad (4)$$

where $T = E_{5,i} + E_{10,i} + E_{20,i}$, and α reflects the response tendency to use the junk category. The probability of placing Pattern i into the junk category can be obtained directly from the following:

$$1 - P_{CS}[A = 5|i] - P_{CS}[A = 10|i] - P_{CS}[A = 20|i] = e^{-\alpha T}.$$

Mixed Prototype Plus CS Model

We have attempted to specify a mixed model that conforms in a general way to the mixed model described by Homa et al. (1981) but that only requires a small number of additional parameters. The basic idea is that if only one prototype has been formed (say, e.g., Prototype A for Category A), then subjects first compare the similarity of the probe to Prototype A. If the similarity is greater than some criterion amount, then the probe is placed in Category A. Otherwise, subjects resort to an exemplar process represented by the CS model. If more than one prototype is formed, subjects first evaluate the similarity of the probe to each of the available prototypes. The probe is placed in Category A if what occurs is the following conjunction of events: (a) The similarity of the probe to Prototype A is greater than the similarity for the remaining prototypes and (b) the similarity of the probe to Prototype A is greater than a criterion amount. Otherwise, subjects resort to an exemplar process represented by the CS model.

Due to the nature of the design employed by Homa et al. (1981), only a small number of parameters need to be added to the CS model to incorporate a prototype-categorization process. Note that the distance between any two patterns from different categories (an average of 10–15 units) was much larger than the distance between a prototype and one of its category members (an average of 5 units with a standard deviation of about .16 units). Based on these figures, the joint probability that (a) an exemplar is most similar to an incorrect prototype and that (b) this similarity is greater than the criterion required for a classification response should be so small that it may be ignored. Thus, we simply employed three prototype parameters μ_5, μ_{10}, and μ_{20}, representing the probability that the subject will correctly categorize an exemplar using a prototype rule when the exemplar belongs to category size 5, 10, or 20, respectively. Based on these assumptions, the probability of categorizing Pattern i as Category A equals the following:

$$P[A|i] = \mu_a + (1 - \mu_a)P_{CS}[A|i], \quad \text{if } i \text{ is a member of } A$$

$$P[A|i] = (1 - \mu_a)P_{CS}[A|i], \quad \text{otherwise,}$$

where $P_{CS}[A|i]$ is the probability of categorizing Pattern i into Category A predicted by the CS model in Equation 4.

This version of the prototype model captures many of the properties described by Homa et al. (1981, p. 422). The probability of using a prototype, μ_a, is assumed to increase with category size for two reasons: One is that the probability of forming a prototype increases with exemplar experience, and the second is that the category boundary

may increase with exemplar experience. Note that the effect of the similarity parameter X decreases as μ_a increases with category size, increasing the Old–New Similarity × Category Size interaction. Classification of the objective prototype can be very low for small category sizes but can be extremely high for large category sizes, by allowing μ_a to vary as category size increases. Finally, because prototype usage is assumed to increase for delayed tests, one would expect μ_a to increase from immediate to delayed tests.

Results

The parameters of the CS and mixed models were estimated separately from four confusion matrices provided by Homa et al. (1981): the results of the immediate and delayed tests for Experiments 1 and 2. A modified Newton–Raphson algorithm (the ZXMIN routine in the IMSL library) was used to estimate parameters that minimized the Pearson chi-square statistic, $\chi^2 = \Sigma(f' - f)^2/f'$, where f and f' represent the observed and predicted frequencies, respectively.[1] The CS model was obtained as a special case of the mixed model by setting $\mu_5 = \mu_{10} = \mu_{20} = 0$.

The estimated parameters for the mixed model are shown in Table 1. The asterisks to the right of some of the parameters indicate parameters that were fixed a priori. (This was for identification purposes, and it does not limit the generality of the model predictions.) The note to Table 1 provides the minimum chi-square statistics, as well as the degrees of freedom (*df*s) for the mixed model. Although it is highly questionable to assume that these chi-square statistics are distributed according to the central chi-square distribution with 52 *df*s, the chi-square statistics for Experiment 1 and for the delayed test of Experiment 2 are less than the critical value for the .05 significance level, indicating adequate fits. The fit to the immediate test of Experiment 2 is much worse, which is perplexing because it was a replication of Experiment 1.

The predicted and observed proportions for each stimulus condition are shown in Table 2. The first column indicates the objective, old–new similarity, where *old* represents an old stimulus, 1 represents a new transfer stimulus with a 1-unit distance between the probe and the closest same-category member, and so on up to *proto*, which represents a prototype pattern. The word *foil* under the first column represents the foils. The second column indicates the category membership of the stimulus. The remaining columns indicate the proportion of times that a given stimulus was placed in each category response. The proportions under the letters *Obs* are the observed proportions, and the proportions under the letters *Mix* are the predicted proportions.

Table 1
Parameter Estimates for the Mixed Model

	Experiment 1		Experiment 2	
Parameter	I	D	I	D
$X(0)$*	1	1	1	1
$X(1)$	0.81	0.55	1.0	0.87
$X(2)$	0.38	0.55	0.76	0.70
$X(3)$	0.16	0.45	0.63	0.56
$X(4)$	0.13	0.23	0.49	0.53
$X(5)$	0.012	0.23	0.32	0.42
$X(p)$	0.04	0.07	0.08	0.08
$S(w)$	0.012	0.016	0.008	0.01
$S(b)$	0.012	0.016	0.005	0.01
$\beta(5)$*	1	1	1	1
$\beta(10)$	0.90	0.88	0.92	1.0
$\beta(20)$	0.80	0.65	0.71	0.80
α	1.63	1.42	3.16	2.83
$\mu(5)$	0.44	0.44	0.0	0.0
$\mu(10)$	0.54	0.52	0.0	0.0
$\mu(20)$	0.71	0.61	0.0	0.0

Note. I = immediate. D = delayed. For Experiment 1, I test: $\chi^2(52) = 60$, D test: $\chi^2(52) = 67$. For Experiment 2, I test: $\chi^2(52) = 98$, D test: $\chi^2(52) = 65$.
* Fixed parameter.

The most crucial question is whether there are substantial differences between the fits of the CS and mixed models. Because the CS model is nested within the mixed model, this reduces to the question of whether the additional three prototype parameters (μ_5, μ_{10}, μ_{20}) produce a substantial improvement in fit. The results of Table 1 provide some

[1] Proportions for new and old patterns were based on 120 observations. Proportions for prototype patterns were based on 48 observations. Proportions for foils were based on 360 observations.

information relevant to this question. Note that the probability of using a prototype was a substantial and increasing function of category size for Experiment 1, but the probability of using a prototype was zero in Experiment 2.

A more rigorous way to compare the CS and mixed models is to perform a chi-square difference test (cf. Bonett & Bentler, 1983). This was performed by calculating the chi-square difference measure, χ^2 (CS vs. mixed) = $\Sigma(f_M - f_{CS})^2/f_{CS}$, where f_M is the frequency predicted by the mixed model, and f_{CS} is the frequency predicted by the CS model. The results of this comparison are presented in Table 3. The rows of Table 3 provide the chi-square difference measures. Again it is highly questionable to assume that these chi-square statistics are distributed according to a central chi-square distribution with 3 *df*s. However, given this assumption, then the deviations between the CS and the mixed models seem to be reliable for Experiment 1, and there is no difference for Experiment 2 because the best-fitting prototype parameters were all zero.

An alternative method for interpreting the difference in model fits is to use a normed fit index (Bonett & Bentler, 1983). This is a descriptive index measuring the relative improvement produced by the prototype parameters, and the interpretation of these measures does not require distribution assumptions. This index relies on the definition of a null or equal probability model, which can be defined as the model that predicts that $P[A|i] = .25$, for all stimulus conditions. The chi-square defined by χ^2 (null vs. mixed) = $\Sigma(f_M - f_N)^2/f_N$ (where f_N is the frequency predicted by the null model) provides a chi-square measure of the total amount predicted by the mixed model. Finally, the ratio χ^2 (CS vs. mixed)/χ^2 (null vs. mixed) gives the relative improvement produced by the additional prototype parameters. For Experiment 1 (both immediate and delayed tests), this ratio equaled .01, whereas for Experiment 2, the ratio was zero.

An attempt was made to isolate one possible cause for the difference in chi-squares between the mixed versus the CS models in Experiment 1. The weakest theoretical assumption of the CS model was the treatment of the junk responses. For the foils, the response distributions are largely determined by the model of the junk response. Thus, it may be of interest to compare the models when the fits are limited to stimuli from Categories 5, 10, or 20 (i.e., the foils are excluded but the junk responses to nonfoils are still included).

The chi-square difference for the immediate test excluding foils was as follows: χ^2 (CS vs. mixed) = 6.1, which is nonsignificant, assuming a central chi-square distribution. Thus, the original chi-square differences between the CS and the mixed models for the immediate test of Experiment 1 seemed to be largely due to responses to foils.

The chi-square difference for the delayed test excluding foils was as follows: χ^2 (CS vs. mixed) = 15, which is significant, assuming a central chi-square distribution. Thus, a moderate amount of the original chi-square difference for the delayed test of Experiment 1 was not due to responses to the foils. In order to further isolate the differences between the CS and the mixed models, an analysis of the absolute deviations between the observed and predicted proportions was computed, excluding foils but including junk responses. This analysis failed to show any consistent differences. In fact, the average absolute deviation for the CS model was exactly equal to the mixed model (.023 for both models, excluding foils but including junk responses).

Several final points concerning the present analyses should be mentioned. First, note that the probability of using a prototype did not increase with delay. Instead, prototype usage tended to decrease in Experiment 1 (no decrease could occur in Experiment 2 because it was already at zero). This result contradicts prototype theories because they generally predict that usage of the prototype should increase with delays between training and test.

Second, the ordering of the bias parameters, (β_5, β_{10}, β_{20}), was consistently negatively correlated with category size across all four confusion matrices. There are two possible explanations for this. First, the guessing strategies may be influenced by the natural tendency for subjects to assign the same number of stimulus presentations to each of the avail-

(*text continued on page 646*)

Table 2
Predictions for the Mixed Model and the Observed Proportions

Old–new distance	Category size	Category response proportion: 5 Obs	5 Mix	10 Obs	10 Mix	20 Obs	20 Mix	J Obs	J Mix
				Experiment 1: Immediate					
Old	5	.84	.83	.03	.04	.04	.07	.08	.06
1	5	.86	.78	.03	.05	.04	.08	.07	.08
2	5	.69	.68	.04	.06	.12	.11	.15	.15
3	5	.60	.58	.08	.07	.10	.13	.22	.22
4	5	.61	.56	.05	.08	.12	.14	.23	.23
5	5	.45	.49	.11	.08	.16	.15	.28	.28
Proto	5	.42	.56	.08	.08	.27	.13	.23	.23
Old	10	.01	.02	.93	.87	.05	.07	.02	.05
1	10	.03	.02	.85	.84	.06	.07	.07	.06
2	10	.03	.03	.75	.75	.08	.10	.13	.13
3	10	.03	.03	.63	.67	.16	.11	.18	.18
4	10	.03	.04	.68	.66	.06	.11	.23	.19
5	10	.03	.04	.60	.60	.15	.13	.22	.23
Proto	10	.00	.03	.73	.71	.17	.10	.10	.16
Old	20	.00	.01	.01	.02	.95	.93	.04	.03
1	20	.00	.01	.02	.03	.92	.92	.07	.04
2	20	.01	.02	.03	.03	.88	.87	.09	.08
3	20	.02	.02	.03	.04	.85	.83	.10	.11
4	20	.01	.02	.05	.04	.78	.82	.16	.12
5	20	.01	.02	.05	.04	.83	.79	.12	.14
Proto	20	.00	.02	.04	.03	.96	.88	.00	.06
Foil	Junk	.10	.08	.17	.15	.28	.27	.46	.49
				Experiment 1: Delayed					
Old	5	.83	.81	.06	.05	.07	.07	.05	.06
1	5	.72	.72	.10	.07	.07	.10	.12	.12
2	5	.77	.72	.07	.07	.08	.10	.09	.12
3	5	.68	.69	.11	.07	.08	.10	.13	.13
4	5	.65	.61	.08	.08	.08	.12	.20	.18
5	5	.59	.61	.12	.08	.11	.12	.18	.18
Proto	5	.58	.62	.15	.08	.06	.12	.21	.18
Old	10	.02	.03	.88	.85	.08	.07	.03	.05
1	10	.07	.03	.78	.78	.09	.09	.07	.10
2	10	.01	.03	.80	.78	.10	.09	.09	.10
3	10	.01	.04	.77	.75	.13	.10	.10	.12
4	10	.05	.04	.68	.69	.13	.11	.13	.16
5	10	.06	.04	.69	.69	.14	.11	.11	.16
Proto	10	.00	.04	.77	.76	.10	.09	.13	.11
Old	20	.01	.03	.04	.05	.88	.88	.08	.04
1	20	.03	.03	.03	.06	.86	.83	.09	.08
2	20	.03	.03	.08	.06	.78	.83	.12	.08
3	20	.00	.03	.05	.06	.85	.81	.10	.09
4	20	.04	.04	.01	.07	.78	.77	.17	.13
5	20	.03	.04	.03	.07	.81	.77	.13	.13
Proto	20	.00	.03	.04	.05	.94	.88	.02	.04
Foil	Junk	.12	.10	.17	.18	.27	.27	.44	.45

Table 2 (*continued*)

Old–new distance	Category size	Category response proportion							
		5		10		20		J	
		Obs	Mix	Obs	Mix	Obs	Mix	Obs	Mix
		Experiment 2: Immediate							
Old	5	.98	.89	.01	.04	.02	.05	.00	.02
1	5	.95	.88	.01	.04	.03	.06	.01	.02
2	5	.83	.83	.03	.05	.05	.07	.09	.05
3	5	.86	.79	.01	.05	.05	.08	.08	.08
4	5	.75	.72	.05	.06	.08	.10	.12	.12
5	5	.52	.60	.09	.08	.17	.12	.23	.20
Proto	5	.63	.55	.10	.08	.17	.13	.10	.24
Old	10	.01	.02	.90	.90	.06	.06	.03	.02
1	10	.03	.02	.87	.89	.07	.06	.04	.02
2	10	.03	.03	.86	.84	.08	.08	.03	.05
3	10	.02	.03	.80	.81	.11	.09	.08	.07
4	10	.03	.04	.75	.75	.10	.10	.13	.11
5	10	.05	.04	.66	.64	.18	.13	.12	.19
Proto	10	.02	.04	.92	.76	.00	.10	.06	.10
Old	20	.00	.03	.04	.05	.94	.91	.02	.01
1	20	.01	.03	.06	.05	.93	.90	.01	.02
2	20	.01	.03	.03	.06	.93	.86	.03	.05
3	20	.04	.04	.06	.07	.82	.82	.08	.07
4	20	.01	.04	.04	.08	.82	.77	.13	.11
5	20	.03	.05	.09	.09	.73	.67	.16	.18
Proto	20	.00	.03	.00	.05	.94	.90	.06	.02
Foil	Junk	.08	.07	.19	.13	.29	.32	.43	.48
		Experiment 2: Delayed							
Old	5	.83	.78	.06	.08	.09	.12	.03	.02
1	5	.74	.75	.10	.08	.13	.13	.04	.03
2	5	.70	.70	.10	.10	.16	.15	.05	.05
3	5	.69	.64	.06	.11	.20	.17	.05	.08
4	5	.57	.63	.16	.11	.17	.18	.10	.09
5	5	.48	.56	.19	.12	.21	.20	.12	.12
Proto	5	.52	.54	.20	.13	.15	.20	.13	.13
Old	10	.04	.04	.85	.82	.10	.12	.00	.02
1	10	.06	.04	.81	.80	.11	.13	.02	.03
2	10	.04	.05	.78	.75	.15	.15	.03	.05
3	10	.04	.05	.74	.70	.13	.17	.10	.08
4	10	.06	.05	.74	.69	.14	.17	.06	.09
5	10	.04	.06	.70	.63	.18	.19	.07	.12
Proto	10	.06	.04	.72	.77	.20	.14	.02	.05
Old	20	.04	.04	.08	.09	.84	.84	.04	.02
1	20	.07	.05	.07	.10	.83	.82	.04	.03
2	20	.04	.06	.07	.11	.81	.78	.09	.05
3	20	.05	.06	.10	.12	.73	.73	.12	.08
4	20	.06	.06	.08	.13	.76	.72	.10	.09
5	20	.05	.07	.11	.14	.70	.67	.14	.12
Proto	20	.00	.03	.04	.07	.96	.89	.00	.01
Foil	Junk	.14	.10	.19	.20	.33	.32	.34	.37

Note. The percentage of correct proportions reported in Table 2 for Experiment 2 (delayed test) differ slightly from those reported by Homa et al. (1981) due to the fact that the data for 1 subject are missing. Obs = observed proportion. Mix = predicted proportion. Old = old stimulus. 1 = new transfer stimulus with a 1-unit distance between the probe and the closest same-category member (2 = 2-unit distance, etc.). Proto = prototype pattern. Junk = junk category.

Table 3
Comparison of Fits for the Null, Complete Set (CS), and Mixed Models

	Experiment 1		Experiment 2	
Measure	I	D	I	D
χ^2 (Null, Mix)	3,440	3,324	4,043	2,946
*df*s	52	52	52	52
χ^2 (CS, Mix)	34	30	0	0
*df*s	3	3	3	3

Note. I = immediate. D = delayed. *df*s = degrees of freedom.

able categories. Parducci (1974) empirically demonstrated this effect with category rating scales. If subjects used the CS rule without bias, then a great disparity in the usage of each category would result. In order to compensate for this disparity, subjects could have adopted response tendencies that tended to equalize the usage of the available categories. More concretely, a subject may say to himself or herself, "Gee, I don't know where to put this pattern. Well, I've been putting too many in Category Size 20, so I'll put this one in Category Size 5."

The other explanation for the negative relation between bias and category size is that the number of category members retrieved from memory and compared with the probe stimulus might have been a negatively accelerated function of the category size, N_a, rather than exactly equal to category size. In other words, $E_{a,i} = X + g(N_a - 1)S_w$, where $g(N_a - 1) < (N_a - 1)$, for $N_a = 10$ or $N_a = 20$. The bias parameters might have compensated for this overly restrictive assumption of the CS model.

Two other trends in parameter values seem interesting. One is that the new–old similarity values shift between the immediate and delayed tests in Experiment 1 but are relatively constant across delays in Experiment 2. Because subjects in the first experiment were tested at both delays but those in the second were tested only once, it is possible that the changes in similarity parameters associated with the first experiment reflect effects of the initial test on performance on the later retention test.

The other notable finding is that the similarity parameter for prototype patterns, which were 4.6 units of distance from old category patterns, did not lie between the value for $X(4)$ and $X(5)$, but rather was substantially smaller in three of the four cases (the immediate test in Experiment 1 was the exception). Thus, performance on the prototype was worse than expected based on old–new similarity, which suggests that the objective metric is not monotonically related to the subjective metric of similarity. One explanation for this is that Homa et al. (1981) created new patterns by moving each vertex of an old pattern by an equal amount. The prototype, however, could differ from old patterns by different amounts for different vertices of the forms. Although the distortion units per vertex were roughly comparable in prototype patterns and new patterns at a distance of 4 or 5 units, the resulting subjective similarities might be greater when distortions are uniformly applied to vertices than when they are variably applied.

Discussion

Exemplar Versus Prototype Models

Our analysis leads us to conclusions somewhat different from those reached by Homa et al. (1981). They criticized the CS model as limited in its ability to account for category size effects and argued for, but did not develop, a mixed prototype and exemplar model. We developed one plausible mixed model but did not find consistent support for it. Although the results of Experiment 1 indicated that the prototype plus CS model provided a modest improvement over a pure CS model, the results of Experiment 2 provided absolutely no evidence for the use of a prototype rule. Also, contrary to the predictions of prototype learning theories, the tendency to use the prototype did not increase with delay between training and testing. It is also worth noting that the mixed model did not provide any improved account of category size effects compared with the pure CS model (both versions, in our opinion, do a quite adequate job). On the other hand, the present results indicate that the CS model is also inadequate because there were substantial improvements produced by the prototype parameters in Experiment 1 (immediate and delayed tests).

Our analyses indicated that the differences were primarily located in the responses to the foils.

One might argue that the prototype model employed in the present analysis was inadequate, and this is why we did not obtain consistent evidence favoring the use of a prototype strategy. In response to this argument, one must consider the fact that the fits of the mixed model were excellent for Experiment 1 (immediate and delayed test) and for Experiment 2 (delayed test). Thus, as far as accounting for the majority of the present results, the mixed model apparently did an adequate job. However, the possibility remains that an alternative version of a mixture model (with a more elaborate prototype rule) will eventually be revealed that can provide adequate explanations for these results. However, at this point, further model testing would be post hoc and probably would not be as beneficial as new research on these questions.

On the other hand, it is interesting that the main advantage of the mixed model over the pure CS model in the first experiment was in accounting for responses to foils. This raises the possibility that the simple algorithm for junk responding in the CS model was at fault and that some alternative version of a pure exemplar model would be adequate. Again, such conjectures are secondary to research that may clarify the discrepancies between Experiments 1 and 2 of Homa et al. (1981).

Finally, it should be emphasized that the systemic deviations from the pure CS model reported by Homa et al. (1981, p. 437) still persist with the mixture proto-plus-exemplar model. More specifically, the proto-plus-exemplar mixed model underestimates the observed category size effect for the prototype probe stimulus and for the new stimuli. Thus, contrary to the conclusion drawn by Homa et al., a prototype-abstraction process does not provide an adequate explanation for the residuals from the pure CS model.

An alternative explanation is to allow the similarity parameters of the CS model to vary across category size. One reason for this is that the degree of learning might have varied across category sizes. For example, the effect of overlearning the small-category-size stimuli might have been to increase the steepness of the generalization gradient from old to new stimuli.

Reanalysis of the FS and SM Models

Equation 2 (with $N = 1$) describes a classification rule that is based on the single pattern that is most similar to the test stimulus. This equation assumes that subjective or encoded similarity maps directly onto the objective similarity measure with essentially no variability. If that were true, then according to the SM model, the classification response should be totally deterministic: The new pattern was always objectively more similar to the specific old one, and therefore, it should always have been assigned to the corresponding category.

It seems more in the spirit of the SM model to assume that the encoded similarity has some variability associated with it and that Equation 2 refers to the relative likelihood that the subjective similarity of the new pattern to its paired (old) training pattern is greater than the subjective similarity to the most similar patterns of contrasting categories. But this does not salvage Equation 2. Once we grant that subjective similarity may map onto objective similarity with some variability, one must recognize that within the category of interest (Category *A*), the training pattern with the greatest (subjective) similarity to the test pattern may not be the one the experimenter has designated as most similar (and represented by X in Equation 2). In general, as category size increases, it will be increasingly more likely that the old pattern with the greatest similarity to the probe will not be the designated one. At the same time, as the number of patterns in contrasting categories increases, it becomes more and more likely that one of them will provide the closest match to the test stimulus.

The implications of this analysis are straightforward. If we allow encoded similarity to vary around objective similarity, then the SM or proximity model will predict Specific, Old–New Similarity × Category Size interaction. That is, as category size increases, the effect of specific, old–new similarity should decrease. One cannot derive precise quantitative predictions without additional assumptions, but this version of the SM model would

predict the observed interaction at least qualitatively.

There is another issue concerning the FS model when $N > 5$. Suppose subjects fix the sample size at $N = 10$, but an exemplar from Category Size 5 is presented. Then, despite the fixed set strategy, the effective sample size will vary with category size when there are only 5 exemplars available to sample. This would force the FS model to predict a Category Size × Old–New Similarity interaction similar to the CS model. A fixed sample model is only possible when $N \leq 5$ in the present study.

Conclusion

An important lesson was gained from this research effort, that is, it is very difficult to intuit the predictions of either exemplar or prototype models. Hintzman and Ludlam (1980) arrived at a similar conclusion when they showed that exemplar models were able to predict the differential forgetting effects and better performance for prototypes than for old training patterns, which were once considered strong evidence for a prototype-classification rule. Considering the results of Homa et al. (1981), it would be difficult to anticipate the form of the Old–New Similarity × Category Size interaction predicted by an exemplar model without a formal analysis. It would also be very difficult to anticipate the contribution of the prototype to this interaction without formal analysis. In general, the use of formal model comparisons rather than intuitions might not only help distinguish exemplar from prototype models, but also might generate new predictions and phenomena that would facilitate understanding of abstraction.

One last point worth noting is that the pure exemplar model and the mixed proto-plus-exemplar model presented here are only special cases of more general theories. The experiments by Homa et al. (1981) do not provide qualitative, parameter-free tests that distinguish between these theories. More specifically, the pure exemplar and the mixed proto-plus-exemplar models both predict Category Size × Old–New Similarity interactions in the same direction. Thus, one is forced to resort to model-fitting methods to evaluate the finer quantitative differences between models. Ideally, future researchers should consider designing parameter-free, qualitative tests of the general theoretical issues.

References

Bonett, D. G., & Bentler, P. M. (1983). Goodness-of-fit procedures for the evaluation and selection of log-linear models. *Psychological Bulletin, 93,* 149–166.

Hintzman, D. L., & Ludlam, G. (1980). Differential forgetting of prototypes and old instances: Simulation by an exemplar based classification model. *Memory & Cognition, 8,* 378–382.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 418–439.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 103–189). New York: Wiley.

Parducci, A. (1974). Contextual effects: A range frequency analysis. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 2, pp. 128–141). New York: Academic Press.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3,* 382–407.

Smith, J. E. K. (1980). Models of identification. In R. S. Nickerson (Ed.), *Attention and performance* (Vol. 8, pp. 129–158). Hillsdale, NJ: Erlbaum.

Townsend, J. T., & Landon, D. E. (1982). An experimental and theoretical investigation of the constant ratio rule and other models of visual letter confusion. *Journal of Mathematical Psychology, 14,* 119–162.

Received April 20, 1983
Revision received April 6, 1984 ■