ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/232562538

# Straight talk about linear separability

Article *in* Journal of Experimental Psychology Learning Memory and Cognition · May 1997

DOI: 10.1037/0278-7393.23.3.659

CITATIONS
105

READS
81

**3 authors**, including:

John Paul Minda
The University of Western Ontario

**35** PUBLICATIONS **1,168** CITATIONS

SEE PROFILE

All content following this page was uploaded by John Paul Minda on 07 June 2014.

The user has requested enhancement of the downloaded file. All in-text references underlined in blue are added to the original document and are linked to publications on ResearchGate, letting you access and read them immediately.

Journal of Experimental Psychology:
Learning, Memory, and Cognition
1997, Vol. 23, No. 3, 659–680

Copyright 1997 by the American Psychological Association, Inc.
0278-7393/97/$3.00

# Straight Talk About Linear Separability

J. David Smith
State University of New York at Buffalo

Morgan J. Murray, Jr.
New School for Social Research

John Paul Minda
State University of New York at Buffalo

It is intuitive that prototypes and additive similarity calculations might underlie human categorization, promoting a special appreciation of linearly separable categories. The failure to document empirically this appreciation has helped focus interest instead on exemplar strategies, multiplicative similarity calculations, and theory-based categorization. However, existing studies have mainly sampled poorly differentiated categories with small exemplar sets. Therefore, the present research repeated existing studies on linear separability, using better differentiated categories better stocked with exemplars. Both the data patterns and modeling suggest that prototypes and a linear separability constraint may have a stronger influence on categorization for these alternative category structures. The information-processing basis for this result is discussed.

## Prototypes, Additive Evidence Rules, and the Linear Separability Constraint

A basic cognitive task is to group objects (foods, predators, mates) into psychological equivalence classes, allowing their similar treatment. An intuitive view of this process is that category members share a family resemblance grounded in common features. According to prototype theories, for example, category members are similar for being variations on a central tendency that participants learn and use in categorization. A generalized prototype principle long anchored the literature on categorization (Homa, Rhoads, & Chambliss, 1979; Homa, Sterling, & Trepel, 1981; Mervis & Rosch, 1981; Posner & Keele, 1968, 1970; Rosch, 1973, 1975; Rosch & Mervis, 1975; Smith, Shoben, & Rips, 1974).

To use a prototype, the categorizer could determine the features a token shares with it and ask whether the combined evidence (similarity) meets some criterion for inclusion in the category. Typical category members would meet this criterion more quickly than atypical category members, producing the well-known typicality effects. Atypical category members could even persistently be categorized into an opposing category if they resembled the opposing prototype.

Categories organized around a prototype will tend to occupy a discrete and coherent region of multidimensional stimulus space, a region that is partitionable away from opposing categories. This coherence can be illustrated by using the stimulus space in Figure 1A. Suppose the lower-left stimuli and the upper-right stimuli are partitioned into categories (Figure 1B). Though neither size nor brightness is diagnostic of category membership, the observer can still combine these independent sources of information, and, using this additive evidence rule, differentiate successfully the larger–darker and smaller–lighter stimuli. In contrast, suppose one category includes the bottom-left and top-right stimuli, whereas the other category includes the top-left stimuli and bottom-right stimuli (Figure 1C). This partition would be difficult to learn using prototypes, because the stimuli do not occupy separable and coherent regions of the stimulus space, and because the prototypes (the categories' central tendencies) are coincident in the space.

Therefore, if prototypes and additive evidence rules do guide human categorization, then category learners will appreciate coherent and separable exemplar pools in stimulus space and may assume that categories are organized in this way. In turn this might constrain the kinds of categories they find natural to learn. Formally, this constraint would imply that humans prefer categories that are linearly separable or separable by a linear discriminant function (Medin & Schaffer, 1978; Sebestyen, 1962). Linearly separable categories (Figure 1B) are those that can be partitioned on the basis of a weighted, additive combination of component information. The geometric sense of this additive evidence rule is that a line through the two-dimensional stimulus space cleanly separates the two exemplar sets. In contrast, the categories of Figure 1C are not linearly separable. There is no way to combine the independent information from size and brightness to unambiguously categorize the stimuli, and there is no line that cleanly separates the exemplar sets.

Isn't it plain that the absence of differentiable prototypes and coherent, separable exemplar pools will defeat learning in the stimulus partition of Figure 1C? After all, in this

J. David Smith, Department of Psychology and Center for Cognitive Science, State University of New York at Buffalo; Morgan J. Murray, Jr., Graduate Faculty, New School for Social Research; John Paul Minda, Department of Psychology, State University of New York at Buffalo.

Correspondence concerning this article should be addressed to J. David Smith, Department of Psychology and Center for Cognitive Science, Park Hall, State University of New York at Buffalo, Buffalo, New York 14260. Electronic mail may be sent via Internet to psysmith@acsu.buffalo.edu.

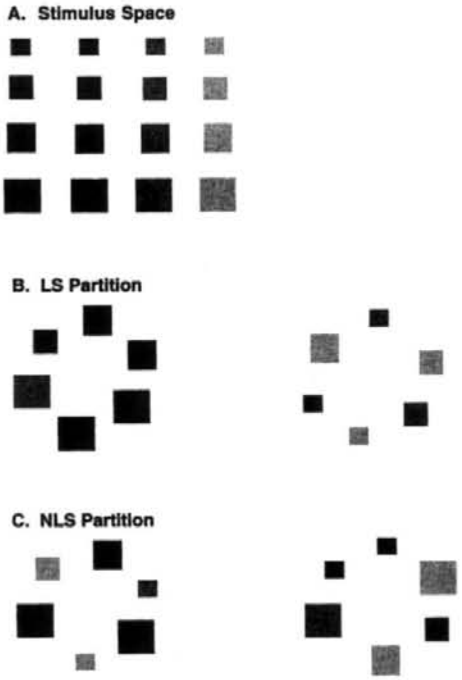

*Figure 1.* A size and brightness stimulus space (A) partitioned into categories that are linearly separable (B) or not (C). LS = linearly separable; NLS = nonlinearly separable.

example the two categories are not well differentiated by basic perceptual similarity. Actually, participants might transcend this difficulty in a variety of ways. They might possess exemplar-based strategies powerful enough to label correctly the stimuli even though they are "disorganized" in multidimensional space. They might code the stimuli more relationally and as configural wholes and note which whole constellations of features deserve which category labels. They might note conceptually that positively correlated size and darkness occupy one category, and negatively correlated size and darkness the other.

These possibilities have been an integral part of the recent categorization literature—for sound empirical reasons (Malt & Smith, 1984; McKinley & Nosofsky, 1995; Medin, Altom, Edelson, & Freko, 1982; Medin & Schaffer, 1978; Medin & Schwanenflugel, 1981; Murphy & Medin, 1985; Nosofsky, 1986, 1987). In fact, participants are capable of learning NLS categories of the kind shown in Figure 1C (Nosofsky, 1987)—they show no evidence of a linear separability constraint that makes only linearly separable (LS) category structures natural and learnable. Lingle, Altom, and Medin (1983, pp. 93–94) even argued that the structures of Figures 1B and 1C might be learned equally easily by a participant who coded stimulus dimensions relationally. Stopping short of this strong claim, an important question remains as to whether, and when, linear separability constrains human categorization (Ashby & Gott, 1988; McKinley & Nosofsky, 1995).

## Evaluating the Linear Separability Constraint

Exploring this question, Medin and Schwanenflugel (1981) noted the dominance of prototype models and additive evidence rules in the literature, the theoretical importance of an LS constraint if it did actually limit the range of natural and learnable category structures, and a surprising empirical silence on this issue. Accordingly, they asked whether nonlinearly separable (NLS) categories were unnatural and poorly learnable for participants. In fact, they asked the question elegantly, for they matched the LS and NLS categories for their within- and between-category similarities (assuming equal salience across dimensions). In contrast, linear separability and stronger similarity relations were confounded in the stimulus sets of Figure 1.

Figure 2 (A–C) shows the structure of three-dimensional LS and NLS categories used by Medin and Schwanenflugel (1981, Experiment 4). The category structures have similar distributions of features with values of one and zero favoring Categories A and B, respectively, always for two of three exemplars in a category. In the LS case, each exemplar has two of three features in common with the category prototypes: 111 and 000. Thus all items can successfully be classified by using an additive evidence rule that compares each token to the prototypes and looks for an evidence score of 2.0. Figure 2–C shows the stimulus space for these categories. One can easily imagine the plane (three minus one dimensions) that partitions the cube (three dimensions) and separates the As and the Bs. These categories are planarly (linearly) separable.

In the NLS case an additive evidence rule will not suffice. For one thing, it will produce categorization errors for Stimuli A1 and B3, which have more features in common with the opposing prototype. For another thing, the NLS categories contain category members that are complements of one another because they have no common features (i.e., A1 and A2 and B2 and B3). These stimulus pairs rule out successful categorization using any linear discriminant function. That is, any weighting of the three independent cues that allows the successful classification of Stimulus A1 (e.g., a very heavy weighting on the first feature) will ensure the misclassification of Stimulus A2. Figure 2C shows that these complementary stimuli are maximally distant in the stimulus space. These categories are not planarly (linearly) separable, for no plane could ever partition the stimulus space and separate the As and Bs.

These two aspects of NLS category structures—exception items and complementary stimulus pairs—are connected. To the extent that one of the complementary stimuli belongs with its category mates (e.g., A2 with A3), the other member of the complementary pair will tend to be an outlier, an exception to the category, and it will seem in an intuitive prototype sense to be "trying to belong" to the opposing category (e.g., A1). If one member of the complementary pair is really prototypical in the category, its complement will be really exceptional in the category.

If humans rigidly depend on prototypes in categorization, then NLS categories might be unlearnable. The exception

**A. LS Categories**

| A | B |
|---|---|
| 011 | 100 |
| 110 | 001 |
| 101 | 010 |

**B. NLS Categories**

| A | B |
|---|---|
| 100 | 000 |
| 011 | 001 |
| 111 | 110 |

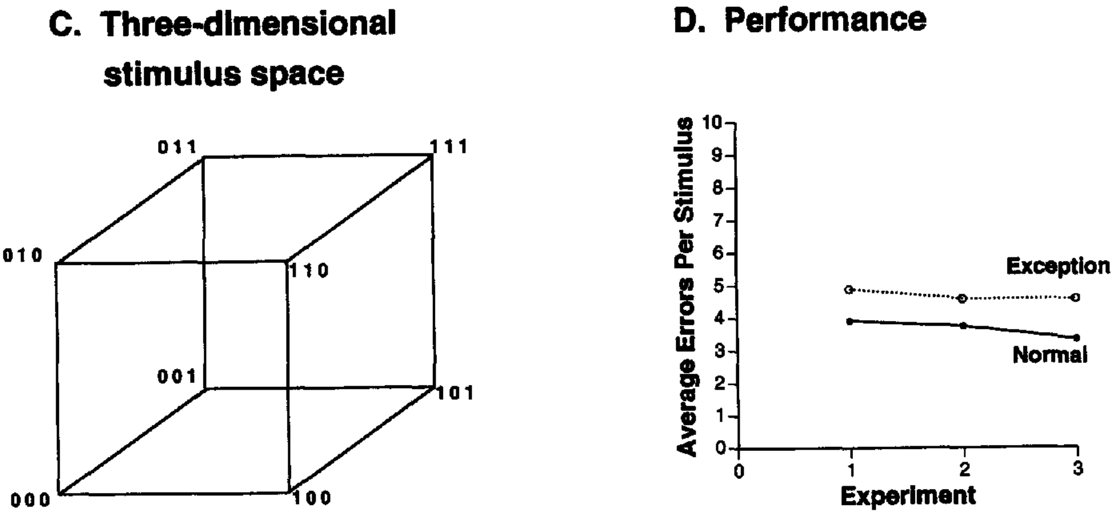

*Figure 2.* LS (A) and NLS (B) categories used in Experiment 4 of Medin and Schwanenflugel (1981). (C) The stimulus space of those categories. (D) Performance on normal and exception items in three NLS category tasks (on the horizontal axis: 1 refers to Medin & Schaffer, 1978, Experiment 1, Stimulus 15 vs. all others; 2 refers to Medin & Schwanenflugel, Experiment 1, NLS Stimulus A1 vs. seven others; and 3 refers to Medin & Schwanenflugel, Experiment 2, NLS Stimuli A1 and B3 vs. six others, generalizations condition). LS = linearly separable; NLS = nonlinearly separable.

items, too similar to the opposing prototype, might lastingly be categorized incorrectly, yielding below-chance performance on them. The complementary pairs of items might lastingly be wrongly placed into opposing categories. Discussing this linear separability constraint, Wattenmaker, Dewey, Murphy, and Medin (1986) stressed that categories must be linearly separable "for a prototype process to work in the sense of accepting all members and rejecting all nonmembers" (p. 160). In fact, any categorization process based on a weighted function of independent properties implies that NLS categories should produce an indefinitely long struggle with exception items (Wattenmaker et al., 1986).

Yet research has not documented this struggle. Medin and coworkers found that the errors made on exception items were only slightly greater than the errors made on normal category members (Figure 2D), and performance on exceptions was better than chance, not below chance as if the stimuli had been assimilated to the opposing prototype. Participants were not just slavishly using a prototype strategy. Instead, they may have been processing whole featural complexes relationally or memorizing individuated exemplars.

This flexibility in processing NLS categories has helped discourage categorization models that emphasize the summed evidence provided by separable and independent perceptual cues and has encouraged an important shift in categorization theory and research. Now the emphasis is on exemplar-based models of categorization, multiplicative calculations of similarity, and conceptually based categories. This shift is important, given a centuries-old conviction by philosophers and psychologists that prototypes and additive evidence rules do matter (e.g., Locke, 1690/1959). The purpose of our article is to encourage more research on when these different categorization strategies will appear and when different categorization models will be appropriate. The existing research on NLS categories has only filled in part of the picture.

## Interpretative Limits on Existing Research

### Small Exemplar Sets

McKinley and Nosofsky (1995) noted that existing research has mainly featured small exemplar pools (about four items per category), whereas prototypes and a linear separability constraint might emerge only for large exemplar pools. In fact, Medin and Schwanenflugel (1981) and Medin, Dewey, and Murphy (1983) worried that their small categories would encourage a paired-associate learning strategy wherein the LS constraint could not show because there were only a few stimulus–category pairings to be learned. Therefore, in additional experiments they included many specific tokens of the few logical types in each category. They still found evidence for exemplar-based strategies and still found no constraint favoring the learning of LS categories. Unfortunately, though, these experiments raised two problematic issues. First, if participants considered the stimuli as abstract types, varying along four dimensions, the effective exemplar pools would still be only about four per category, and not the number of specific tokens (e.g., with different instantiations of blonde hair or shirt color). Second, these categories were difficult for participants to learn—38%, 66%, or 72% of participants failed to reach even a modest learning criterion (Medin et al., 1983, last-name-infinite condition; Medin & Schwanenflugel, 1981, Experiments 3 and 4).

### Learning Difficulties

In studies beyond those just mentioned, 30%, 40%, or 60% of participants failed ever to meet the preset learning criterion (four experiments in Medin & Schaffer, 1978; Medin & Schwanenflugel, 1981; Medin & Smith, 1981). In contrast, participants in other studies have met a criterion of 36 errorless trials (Hartley & Homa, 1981), 70 trials (Homa et al., 1981), or even 90 trials (Homa et al., 1979). By another measure, participants in the existing experiments have an asymptotic performance of about 80%, even after receiving 108–288 potential training trials (Medin & Schaffer, 1978, Experiment 2; Medin & Schwanenflugel, 1981, Experiment 3). In contrast, Kemler Nelson (1984) and Smith and Shapiro (1989) used categories that let participants exceed 90% accuracy with only 24 training trials. In Homa et al. (1979) and Hartley and Homa (1981) participants achieved almost perfect learning. Indeed participants often exceed 80% correct after a week's delay, when given only 6 s to study the eight stimuli in two categories, or under incidental conditions (Hartley & Homa, 1981; Homa et al., 1979; Homa et al., 1981; Kemler Nelson, 1984; Smith & Shapiro, 1989). Thus, research must still document the character of categorization strategies when the categories are "easier" to learn.

### Impoverished Category Structure

Existing LS–NLS comparisons have typically featured impoverished category structures. Structural ratio, which can be defined as the ratio of the similarity of exemplars within categories to the similarity of exemplars between categories, is a useful way to demonstrate this fact about existing LS–NLS comparisons (see also Homa et al., 1979, pp. 13–14). Figure 3A shows a space containing 700 hypothetical category structures. The stimuli for these categories were constructed by using six binary dimensions. There were seven exemplars in each category, derived from the prototypes 0 0 0 0 0 0 and 1 1 1 1 1 1 for the two categories. To make this figure, the 42 stimuli that had three or more features in common with a prototype were divided into three classes: Class 1 (those with six or five common features), Class 2 (those with four common features), and Class 3 (those with three common features). Then 35 selection procedures were generated as follows: 6 1 0, 6 0 1, 5 2 0, 5 1 1, 5 0 2, . . . 0 0 7. Thus, the first 20 stimulus sets contained, in each category, 6 Class 1 stimuli and 1 Class 2 stimulus. The last set of 20 stimulus sets contained 7 Class 3 stimuli in each category. All repeats of items between or within categories were ruled out. To be fair to the two influential ideas about similarity calculations, the index of category structure has been calculated by using both additive and multiplicative similarity rules. The correlation between the two indices was .97, making the kind of similarity calculation one adopts for the present purpose a matter of indifference.

Figure 3B shows in the same space the 12 category structures that have been used for six existing comparisons between LS and NLS categories. These comparisons can be found in Medin and Schwanenflugel (1981, Experiments 1–4), Kemler Nelson (1984, Experiment 3), and Wattenmaker et al. (1986, Experiment 1). There appear to be fewer than 12 data points because some category structures have been repeated. All of these comparisons have used poorly differentiated categories by either metric. This is a natural outcome of the careful stimulus control these experiments require. Yet this poor differentiation means that existing category structures contain highly similar exemplars that participants are very likely to confuse; indeed this has seemed a virtue of the experiments (Medin & Schwanenflugel, 1981, p. 361). Sometimes there are more close matches between exemplars across categories than within categories (Medin & Schwanenflugel, 1981, p. 358). Also in these categories, the prototypes hardly stand out as special, even when they are presented, and individual features are poorly predictive of category membership (50%–75%). Possibly these "categorization" tasks sometimes degenerate into identification tasks, in which participants rotely associate whole instances and their labels but have no sense of organized categories as they apply the labels. Medin and Schwanenflugel (1981, p. 365) entertained just this worry about their specialized categorization tasks. Though humans may depend on exemplar encoding, not prototypes and linear separability, given impoverished category structure, it is still possible that prototypes and linear separability become important given well-differentiated categories (Medin & Schwanenflugel, 1981; Medin & Smith, 1981, pp. 171–172). (One can explore more fully the possible relations between exemplar memorization and categorization in Shepard, Hovland, & Jenkins, 1961; Medin & Schwanen-

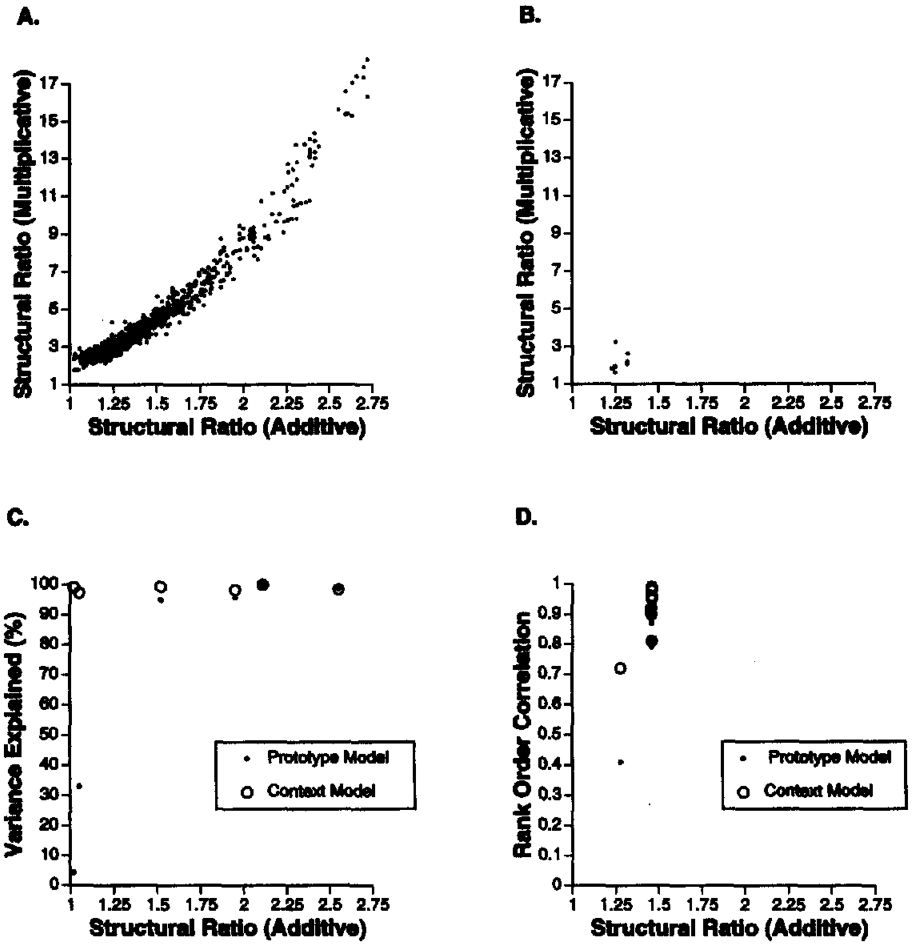

*Figure 3.* (A) Seven hundred hypothetical category structures illustrating the range of category differentiation that is easily available. (B) The level of category differentiation provided in six existing comparisons between linearly separable and nonlinearly separable categories. (C) The fit of two categorization models to humans' performance when they are given categories of different structural ratios (see Nosofsky, 1987). (D) The fit of two categorization models to humans' performance when they are given categories of different structural ratios (see Medin & Schaffer, 1978; Medin & Smith, 1981).

flugel, 1981, pp. 361–362; Nosofsky, 1984, 1986, 1987; and Smith, 1989.)

## Category Differentiation and the Fit of Models

Some evidence hints at a link between good category structure and the use of prototypes in categorization. For example, Nosofsky (1987) gave his participants six different category structures, which varied in structural ratio from 1.17 to 2.39, calculated by comparing the average Euclidean distance among stimuli between categories and within categories. Nosofsky fit a prototype model and an influential exemplar-based alternative to participants' performance with each stimulus set. Where categories are well differentiated, the models perform equally. For poorly differentiated categories, prototype models do not suit at all (Figure 3C).

Medin and Schaffer (1978) and Medin and Smith (1981) also fit exemplar-based and prototype-based models to categorization performance (Figure 3D). In conditions of low-to-moderate category structure (structural ratio = 1.46), both models performed equivalently when different indices of observed and predicted performance were correlated. One condition had weaker category structure (structural ratio = 1.28), and it produced the one decisive failure of the prototype model. Apparently, the prototype model gains rapidly over structural ratios of 1.2 to 1.5, for two- or four-dimensional stimulus sets, integral or separable dimensions, and continuously varying or binary dimensions.

There are a variety of possible perspectives on these graphs. One can point to the exemplar model's advantage for poorly structured categories, or one can worry that that model fits best only when prototype strategies are undermined and exemplar strategies are mandatory. One can note the exemplar model's greater staying power, for it handles "bad" and "good" category structures, or one can note that prototype models achieve parity just where they should and only where they could—where robust, differentiable prototypes are on the scene. Or, one can downplay both good fits for well-differentiated categories because formally the models must fit well and must start to converge for the homogeneously excellent performances supported by higher levels of category structure and predicted by both models.

Ideally, though, the question could be a psychological one, not a formal one. Figures 3C and 3D suggest the possibility of an interesting psychological transition that occurs as structural ratio improves, gradually suggesting prototype strategies, and gradually making them useful. Of course it will require diagnostic category structures to show this transition if it exists. Then one may be able to tell what strategies emerge for different kinds of categories and what models are appropriate for describing different performances. In our view, this remains an important unknown in the existing literature.

## Experiment 1

Existing research on NLS categories has sampled an exemplar-poor, poorly differentiated, and poorly learnable region of the universe of category structures. Consequently, the inference remains only specific and limited that humans use exemplar-based strategies and are free from a constraint favoring linear separability. Accordingly, in Experiment 1 we duplicated existing comparisons between LS and NLS categories, using better differentiated, more exemplar-rich categories. We explored the possibility that these alternative category structures will encourage the use of prototype strategies and will subject categorizers to a linear separability constraint.

### Method

*Participants.* Sixty-four members of a university community were paid to participate in this study.

*Stimuli and category structures.* The crucial aspect of the design was to use six-dimensional stimuli instead of the usual three or four dimensions. This allowed larger exemplar pools and better differentiated categories than were previously used.

The stimuli were pronounceable six-letter nonsense words with the following pattern of consonants (Cs) and vowels (Vs): CVCVCV (see also Smith & Shapiro, 1989; Smith, Tracy, & Murray, 1993; Whittlesea, 1987; Whittlesea, Brooks, & Westcott, 1994). Stimulus generation began with the creation of two prototype pairs (*bunoli–kypera* and *girupo–letany*). The first and second members of each pair were designated as stimuli 0 0 0 0 0 0 and 1 1 1 1 1 1, respectively. These prototypes were created randomly but with several constraints to ensure the pronounceability of all stimuli, the orthographic appropriateness of all stimuli, the identical syllabification of all stimuli, and the roughly equal use of all vowels. For example, *q* (which is followed by two vowels), *c* (which changes sounds depending on the vowel following), and *e* in final position (which often changes syllabification) were disallowed. Each prototype pair contained six different vowels and six different consonants.

The Appendix (top) shows the structure of the LS and NLS stimulus sets with poorer category structure that should present more performance difficulties. Each LS category contained seven stimuli that shared four features with the category prototype. There were no exception items. A prototype strategy, using an additive rule that summed across independent attributes, would allow perfect categorization. Each NLS category contained three stimuli with five typical features, three stimuli with four typical features, and one exception that had five atypical features. The similarity relations in the NLS categories were heterogeneous, but the cluster of highly similar instances combined with the exceptions balanced overall similarity at the level for the LS categories. The complement of the exception also appeared (Stimuli A2 and A7 and Stimuli B3 and B7), ruling out an additive categorization rule and guaranteeing NLS categories.

These general category structures were predetermined to allow the generation of well-matched LS and NLS stimulus sets. After these assignments, hundreds of possible stimulus sets were computer generated and screened for LS and NLS category structures that matched in several ways. The two stimulus sets had identical exemplar–exemplar similarity, both within category (an average of 3.47 shared features) and between category (2.69 shared features), and identical exemplar–prototype similarity, both within category (4.00 features) and across category (2.00 features). By either standard, the two stimulus sets had identical structural ratios (1.29 when the exemplar–exemplar similarities were used) and structural ratios like those in previous LS–NLS comparisons. These calculations assumed additive similarity calculations and equal salience for all of the features.

In addition, the LS and NLS stimulus sets were matched in the overall informativeness of all attributes. For the LS categories, Group A stimuli took the typical value—zero—four, five, four, six, five, and four times for attributes 1–6, and Group B stimuli took the typical value—one—four, four, six, five, four, and five times for attributes 1–6. These numbers were identical for the NLS categories. Thus, there were no criterial attributes at work in these stimulus sets, and any single-letter strategy should have been identically salient and viable in both stimulus sets.

The category structures were not matched, however, if one considers the number of highly similar pairs of exemplars. Neither LS category had any highly similar exemplar pairs (e.g., five shared features), the NLS categories featured seven pairs. As Medin and others have noted, this might grant the NLS categories a learning advantage if exemplar-based strategies dominated categorization and highly similar pairs of exemplars retrieve each other and afford easy categorization.

The six dimensions of the stimulus array were further exploited to create matched LS and NLS category structures with better differentiated categories that participants might find easier to learn. The structure of these LS and NLS categories is also shown in the Appendix. The LS categories contained one prototype, two stimuli with five typical features, and four stimuli with four typical features. There were no exception items; an additive rule would correctly categorize all stimuli. Again the LS similarity relationships were fairly homogeneous but weakened compared with the cluster of highly similar exemplars in the matched NLS categories.

The NLS categories contained one prototype, five stimuli with five features in common with the prototype, and one exception item that shared five features in common with the opposing prototype. The cluster of highly similar instances combined with the exception item balanced overall similarity at the level for the LS task. The complement of the exception stimuli also appeared (Stimuli A5 and A7 and B5 and B7), also ruling out an additive decision rule.

These category structures allowed well-matched LS and NLS stimulus sets that were chosen from among hundreds of computer-generated sets. The two category structures had identical exemplar–exemplar similarity, both within category (3.88 features) and between category (2.12 features), and identical exemplar–prototype similarity, both within category (4.57 features) and between category (1.43 features). By either standard, the two category structures had identical structural ratios (1.83 when the exemplar–exemplar similarities were used) and more category structure than the poorly differentiated stimulus sets.

In addition, LS and NLS categories were matched in the overall informativeness of all attributes. For the LS categories, Group A stimuli took the typical value—zero—five, five, five, six, six, and five times for attributes 1–6, and Group B stimuli took the typical value—one—five, five, five, six, six, and five times for attributes 1–6. These numbers were identical for the NLS categories. Again there were no criterial attributes available in either of these stimulus sets, and any single-letter strategy should have been identically salient and viable in both of them.

The categories were not matched, however, if one considers the number of highly similar pairs of exemplars. Once again the NLS categories offered more highly similar exemplar pairs and might support best exemplar-based strategies in categorization. This could give the NLS categories a learning advantage if exemplar-based strategies dominated the processes of categorization.

For both the poorly structured and moderately structured categories, two LS and two NLS tasks were constructed, one using each prototype pair. Each task contained 196 trials (14 blocks) numbered consecutively without visible breaks. Each block contained one randomly generated run through the seven stimuli in each category. The stimuli were arranged into four random orders for different groups of participants.

*Procedure.* Participants revealed each successive trial through a window cut in white paper. This ensured trial-by-trial presentation and prevented scanning forward and backward along the page. Participants classified each stimulus as an A or B on a separate, numbered answer sheet. The answer to each trial was printed below it, invisible while the participant responded but available afterward as feedback. Participants responded at their own pace—the experimental session took approximately 1 hr. Participants were randomly assigned to LS and NLS conditions and to the particular prototype pair and order of stimulus presentation they would receive.

The instructions for the categorization task were as follows:

> In this experiment you will see nonsense words, each six letters long, which can be classified as Group A or Group B words. Your task in this experiment is to look closely at each word as it appears and figure out how you can tell whether the word belongs to Group A or Group B. At first the task will be quite difficult, but with time, and by studying the words carefully, you should be able to answer correctly.

The instructions included a description of how to use the paper window and how to reveal the correct answer to obtain feedback on one's performance. Participants were told to "study the word and the answer together to try and gain more information about how the words can be classified."

### Results

These categories, even those with moderate category structure, were difficult to learn. Participants were 67% correct overall. Possibly the present manipulations to increase category structure still did not increase structural ratio enough to create easily learnable categories. After all, structural ratios of 1.83 are middling in comparison to the sweep of category structures shown in Figure 3A. Another possibility is that the larger exemplar pools are undermining exemplar memorization, counteracting the benefits of increased structural ratio.

The data were analyzed by using a four-way analysis of variance (ANOVA) with low–moderate structural ratio, LS–NLS, and Prototype 1–2 as between-subjects variables and trial block as a within-subject variable. Two effects were reliable (Figure 4A). First, despite the difficulty of both category structures, significant learning occurred across trial blocks, $F(13, 728) = 14.41, p < .05, MSE = 0.194$. Second, performance was better for the categories with moderate structural ratio, $F(1, 56) = 13.28, p < .05, MSE = 1.819$. Overall accuracy rates were 62% for the poorly structured categories and 71% for the moderately structured categories. This held for the LS categories considered alone (62% vs. 68%) and for the NLS categories considered alone (62% vs. 74%). Final accuracy rates were also higher for the moderately structured categories (70% vs. 80%). This result also held for the LS categories (67% vs. 80%) and the NLS categories (72% vs. 79%). Moreover, 2 participants given the poorly structured categories achieved an errorless block of 14 trials, compared with 8 participants given the moderately structured categories. The success of the manipulation to increase category differentiation reinforces the idea that structural ratio often predicts learnability in a categorization task—both prototype-based and exemplar-based models of categorization comfortably predict this result. A structural ratio of about 1.3, seen in previous LS–NLS comparisons and in the present poorly differentiated category structures, often makes category learning very difficult.

A third effect did not obtain, there was no performance advantage for the LS categories over the NLS categories (Figure 4B). Performance was equivalent throughout learning, $F(1, 56) = 1.48$, *ns*, $MSE = 1.819$. Thus, a constraint favoring linear separability was not evident in overall percentage correct.

However, overall performance is not generally suitable for indicating such a constraint. The problem lies in the different structure of LS and NLS categories. The NLS categories contain a cluster of similar items and one exception item. To achieve equivalent category structure overall, the homogeneous similarity relationships in the LS categories are deliberately weakened. Suppose that this weakening creates performance on 14 LS items of 80%; whereas 12 normal NLS items are performed at 90%, and 2 exception NLS items are performed at 20%. Overall performance is equated. Nonetheless, something has gone badly wrong in the NLS condition. Participants have failed to learn the exception items and therefore have failed to learn the NLS categories in the sense of including category members and excluding nonmembers. One might rather say that participants have insisted on reallocating the exception items into the wrong categories and have constructed good LS categories for themselves. In such a case, a strong linear separability constraint would have been at work in participants' processing, but it would be invisible in overall performance.

Constraint here simply means that participants' decision boundaries are too rigidly "straight" to bend across the stimulus space and place the exception items in the correct categories. The result is systematic errors on the exception items, their effective reallocation into the opposing categories, and the de facto construction of good LS categories by

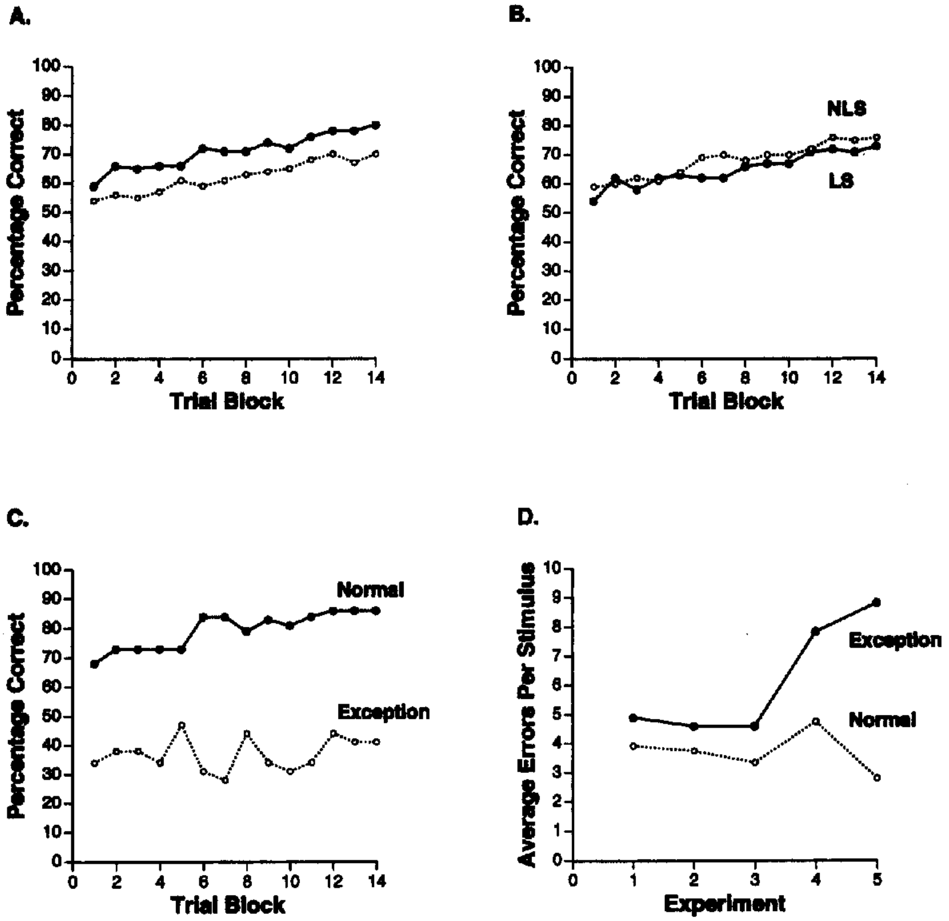

*Figure 4.* (A) Performance on the poorly and moderately differentiated categories by trial block in Experiment 1. (B) Performance on the linearly separable (LS) and nonlinearly separable (NLS) categories by trial block. (C) Performance on normal and exception items by trial block, for all participants in the moderately differentiated NLS condition. (D) Performance on normal and exception items in five NLS category tasks (on the horizontal axis: 1 refers to Medin & Schaffer, 1978, Experiment 1, Stimulus 15 vs. all others; 2 refers to Medin & Schwanenflugel, Experiment 1, NLS Stimulus A1 vs. 7 others; 3 refers to Medin & Schwanenflugel, Experiment 2, NLS Stimuli A1 and B3 vs. 6 others, generalizations condition; 4 refers to the poorly differentiated NLS condition of Experiment 1; and 5 refers to the moderately differentiated NLS condition of Experiment 1).

participants. This constraint is not logically or necessarily linked to any particular processing assumption. The facts of systematic errors, reallocation, and de facto LS categories would still be true, the LS constraint would still hold sway, whether participants were processing the stimuli using prototypes, exemplars, rules, or anything else. However, the literature has traditionally made a strong link between the LS constraint and prototype processing, to which we return (Lingle et al., 1984; Medin & Schwanenflugel, 1981; Murphy & Medin, 1985; Wattenmaker et al., 1986).

The LS constraint was operative in the present data, especially for the moderately differentiated NLS categories (Figure 4C). To assess this effect, we entered the NLS data into a three-way ANOVA with exception–normal items and trial block as within-subject variables and low–moderate structural ratio as a between-subjects variable. The exception items were performed much worse than the normal items, $F(1, 30) = 56.82, p < .05, MSE = 1.409$. In fact, the exception items were performed far below chance, meaning that despite feedback, participants consistently placed the exception items into the wrong category ($z > 2.58, p < .05$), even in the final block.

Moreover, performance had an opposite dynamic between the normal and exception items. There was a significant interaction in the ANOVA between low–moderate structural ratio and exception–normal item performance, $F(1, 30) = 5.74, p < .05, MSE = 1.409$. From the poorly structured to the moderately structured categories, participants became more accurate on the normal items (66% vs. 80%), but less accurate on the exception items (44% vs. 37%).

Recall that Figure 2D showed that the exception items were performed about the same as normal items in poorly differentiated category structures. Figure 4D includes those three studies in a broader perspective, including the results from the present poorly structured and moderately structured NLS categories. In the present case, with larger exemplar sets and better differentiated categories, performance becomes far more heterogeneous because performance on normal items and exception items diverges.

Of course, prototype models account easily for this divergence. Given assimilation to prototypes, participants should insist on forming good LS categories for themselves and should show poor exception–item performance given this reorganization. In fact, this is just what Wattenmaker et al. (1986) predicted would occur when participants rigidly apply prototype representations to NLS categories: They should lastingly place exception items into the wrong categories and thereby reassert the linear separability that a prototype strategy demands (Medin & Schwanenflugel, 1981; Murphy & Medin, 1985). Thus the present results could suggest the interesting possibility that prototype strategies are emerging given the category structures of Experiment 1.

However, formal modeling offers a crucial additional perspective on the NLS data pattern, for it is possible that exemplar models might also predict this reallocation of the exception items and might also predict that participants will operate under a linear separability constraint given the present category structures. We take this possibility seriously, despite the frequent claim in the literature that exemplar processing is a primary strategy by which participants can transcend the LS constraint enforced by prototype models (Lingle et al., 1984; Medin & Schwanenflugel, 1981; Wattenmaker et al., 1986). Accordingly, following Medin and coworkers, we contrasted an exemplar model featuring multiplicative similarity calculations and a prototype model featuring additive similarity calculations. The contrast between these two models has been critical in many formal studies of perceptual categorization. (We note that other models, such as exemplar models with additive similarity calculations and prototype models with multiplicative similarity calculations, are logically possible and potentially interesting too; see Nosofsky, 1992.)

In making the same contrast here, one interesting possibility is that the prototype model might make a better showing in the category structures of Experiment 1 than it has in previous comparisons involving exemplar-sparse, poorly differentiated categories. Another interesting possibility is that the exemplar model might start to fail in fitting some of the participants' data.

## *The Parity Between Models of Categorization*

*Procedure for fitting and testing models.* In evaluating the exemplar model, we focus on the original context model of Medin and Schaffer (1978), preserving the connection to their work and the access to important intuitions. Nosofsky (1984, 1986; see also McKinley & Nosofsky, 1995) provided more formal and general treatments of the context model.

In the exemplar model, the to-be-classified item in the present tasks would be compared with the seven A exemplars (including itself if it is an A), and with the seven B exemplars (including itself if it is a B), yielding an overall similarity index of the item to Category A members and Category B members. Dividing overall A similarity by the sum of overall A and B similarity would yield the probability of an A categorization. Crucially, each comparison between a stored trace and the item would calculate similarity multiplicatively. To use a simplified example, the trace and the item would be compared along each of six dimensions, and the dimensional comparison would yield a 1.0 for a matching feature, .1 for a mismatching salient feature, and .9 for a mismatching nonsalient feature. The similarity between the trace and the item would be the product of these six dimensional weights—1.0 for identity, but .1 for a single, salient difference. One sees from this simple example that similarity decays very rapidly in the exemplar model as featural dissimilarities accumulate; indeed, in some configurations (i.e., for some sets of similarity weights) the exemplar model strongly emphasizes exact identity.

In fitting the exemplar model to target performances, we sampled 46,656 sets of weights in a grid search of the parameter space. Each dimensional weight was varied from .00 to 1.00 in .2 intervals, in nested fashion. For each set of weights we calculated predicted categorization probabilities for the 14 stimuli in a stimulus set, and we compared these predicted performances to the target performance. The measure of fit was the sum of the squared deviations between observed and predicted performance, and in the course of a run that distance was minimized.

Extensive additional simulations confirmed that these grid searches were finding nearly the best fits available in the parameter space. In these simulations, the parameter space was seeded with six different starting configurations (e.g., the best parameters from the grid search, parameters indicating completely divided attention, and parameters indicating complete attention to one or two randomly chosen attributes), and a fine grained, hill-climbing algorithm sought the best fitting state from there. Subsequently, we took the best fit achieved by any of the six hill climbs and subtracted this from the fit achieved by the best node of the grid. Across the 48 individual participants who are the primary focus of the modeling in this article, and who were modeled individually with a grid search and with six hill-climbing seeds, the average improvement in fit was negligible—.006 for the prototype model and .006 for the exemplar model. Given that result, the grid searches have the virtue of being comprehensive and systematic without the need for specifying starting configurations that could favor one model over another. In other cases, though, hill-climbing algorithms may be the preferred alternative.

To evaluate the prototype model, we supposed that each to-be-categorized item would be compared with the category prototype along the six independent dimensions, using additive similarity calculations. Mismatching features contribute zero similarity; matching features contribute the amount of their dimension's weight. The six dimensional weights have typically been required to sum to 1.0. In the simplest case the prototype–item similarity is taken to be the probability of a correct categorization, and its complement the probability of an error. (The prototype's self-similarity is 1.0.)

As is customary, though, an additional free parameter was added to the prototype model because of the fixed sum of the dimensional weights. In our case, this additional guessing parameter captured participants' error proneness with these difficult category structures. Without the guessing parameter, for example, the prototype (with perfect self-similarity) would always have perfect predicted performance. Thus it was assumed that some proportion of the time G participants simply guessed A or B haphazardly. On the other 1-G proportion of trials with that stimulus the participants used prototype similarity as already described (see also Medin & Smith, 1981).

To analyze the prototype model's behavior, a grid search instantiated all of the allowable configurations of the six dimensional weights when they varied from 0 to 1.0 in .1 increments. There were 3,003 configurations that preserved the sum of 1.0. In association with a guessing parameter that also varied from 0 to 1.0, the grid search evaluated 33,033 parameter settings. Once again the sum of the squared deviations between observed and predicted categorization performance was minimized.

Extensive additional simulations investigated whether tending to one inequity (by giving the prototype model an equivalent number of free parameters) created another inequity (by not allowing the exemplar model to incorporate guessing). In these simulations, the parameter space was again seeded with six different starting configurations, and a fine grained hill-climbing algorithm sought the best fitting state of the exemplar model from there. This time, though, guessing was included as a parameter in the exemplar model's search space. Subsequently, we found the best fit achieved by any of each participant's six hill climbs (with guessing available) and subtracted this from the fit achieved by the best (nonguessing) node of the grid. Across the 48 individual participants who are the primary focus of the modeling in this article, and who were modeled individually with a grid search and with six hill-climbing seeds, the average improvement in fit was negligible—.009. Given this result, relying on the nonguessing exemplar model has the advantage of equating the two models for free parameters and of reproducing closely the exemplar model of Medin's original research.

*Fits to individual participants' data.* As a first step, we ran the prototype and exemplar models against each participant's performance in the moderately differentiated NLS condition. The mean fits averaged over participants, .2048 and .2340, were the same. For 8 participants the prototype model fit performance better; for 8 participants the exemplar model fared better. Apparently, the moderately differentiated NLS categories produced an equivalence between models—possibly this is a better showing than the prototype model has achieved in the past.

*The prototype subgroup.* The 8 participants who were fit better by the prototype model averaged 92% correct, 78% correct, and 23% correct on the prototypes, normal items with five typical features, and exception items, respectively. On average, the prototype model fit these participants' data better than did the exemplar model (.1466 and .3154, respectively).

Figures 5A and 5B show the average performance of the 8 prototype participants, compared with the average of the eight predicted performance profiles that fit each participant's data best. The prototype model consistently grants the prototypes (Stimuli 1 and 8) their observed advantage and simultaneously allows dismal performance on the exception items (Stimuli 7 and 14). It heterogenizes performance appropriately. The exemplar model consistently captures these participants' performances less well because it persistently underpredicts prototype performance and overpredicts exception–item performance. Even its best fitting configurations homogenize performance too much for what participants actually do.

Of course one might worry that these 8 participants were really exemplar-based categorizers, until a handful of chance occurrences changed the surface appearance of their performance. To evaluate this possibility, we asked how likely it was that these 8 participants were chance variants on the configuration of the exemplar model that fit best their composite performance. That is, we found the 14 predicted categorization probabilities of that best fitting configuration and created 500,000 exemplar-based categorizers that performed 196 trials in our task, performing according to those predicted probabilities, but probabilistically (i.e., if the predicted probability was 83% of an A response, on any given trial the probability was still 17% of a B response). Chance creates only 2% of the time an observer who performs as well as the subgroup on the prototypes, but as poorly on the exceptions. That half of the 16 participants would aggregate to this performance pattern by chance alone is unlikely indeed. These participants are not a statistical fluctuation around even the configuration of the exemplar model that fits their performance best.

*The exemplar subgroup.* The 8 participants who were fit better by the exemplar model averaged 81% correct, 79% correct, and 51% correct on prototypes, normal items, and exceptions, respectively—this subgroup included many of the best performers on the exception items. On average, the exemplar model fit these participants' data better than did the prototype model (.1536 and .2630, respectively). Figures 5C and 5D compare the averaged observed performances and the averaged best fitting performance profiles. Only the exemplar model fits these data well. The prototype model fits poorly partly because it seriously underpredicts exception–item performance.

Using the statistical procedures already described, we asked whether these participants were really prototype categorizers veiled by a few chance events (e.g., a few unlucky or lucky guesses on the prototypes or exceptions, respectively). Starting with the prototype-based performance pattern that fit best the composite performance of the exemplar subgroup, chance creates only 2% of the time a pattern combining prototype performance below 81% and exception–item performance above 51%. That half of the sample would aggregate to this performance pattern by chance alone is quite unlikely. The exemplar participants are

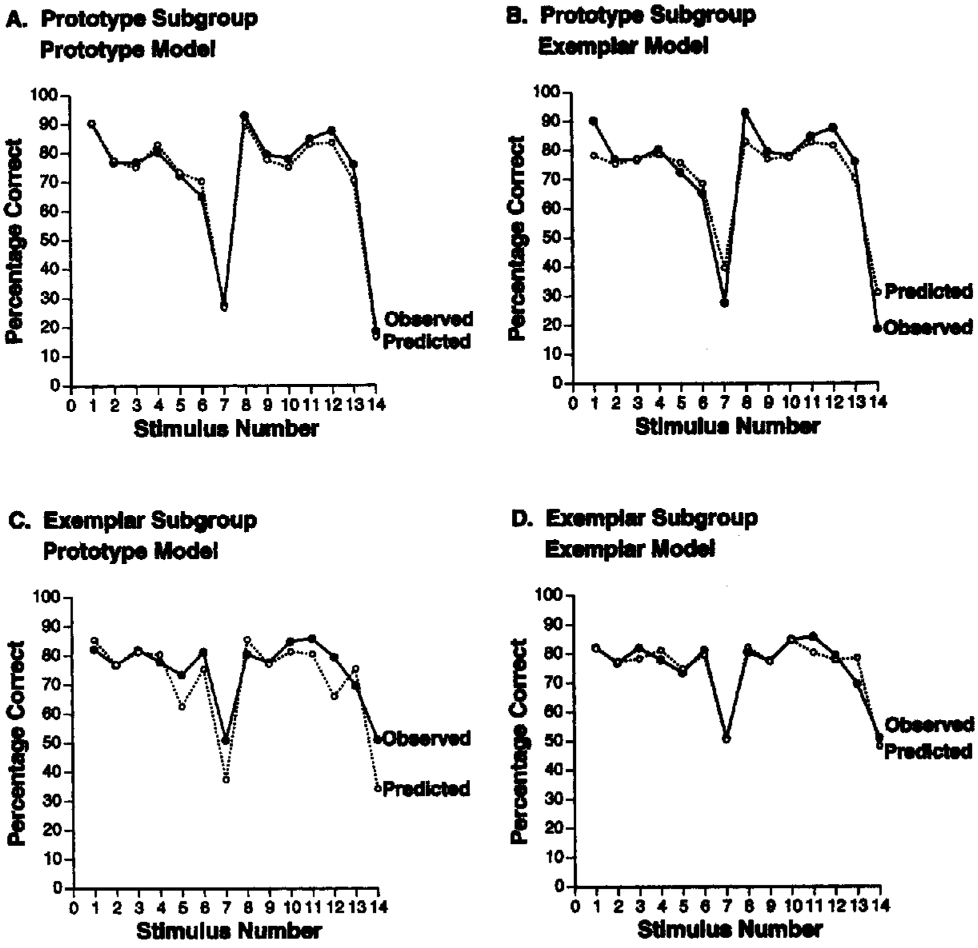

*Figure 5.* (A) The fit of the prototype model to the performances of the prototype participants in Experiment 1. The solid line shows the average of the observed performance profiles produced by the 8 prototype participants. Each participant's performance was then modeled individually, and the dashed line shows the average of the best fitting predicted performance profiles. (B) The fit of the exemplar model to the prototype participants' performances, graphed as in Figure 5A. (C) The fit of the prototype model to the performances of the exemplar participants in Experiment 1. (D) The fit of the exemplar model to the performances of the exemplar participants.

not just statistical fluctuations of even the configuration of the prototype model that fits their performance best.

This subgroup analysis, like the analysis by individual participants, revealed a parity between the prototype and exemplar models—this parity has not been the characteristic claim in the literature.

### *The Problem With Aggregating Data*

We also proceeded in the manner of former studies, by fitting both models to the aggregate performance of the 16 participants. The exemplar model fit better the aggregated data (.0145) than did the prototype model (.0336). Sufficing with this analysis, as others have, we could have concluded that the context model enjoyed its usual advantage. Thus, it is important to document and analyze carefully this problem with modeling aggregated data.

To do so, we produced simulated samples of participants who were known to represent a precise balance between prototype and exemplar strategies. Prototype participants were given a random set of weights from among the 3,003 configurations that had a zero rate of guessing. These weights determined their 14 idealized categorization probabilities. Then each "participant" performed 196 trials in the moderately differentiated NLS condition, with chance allowed to disturb these categorization probabilities around their ideal values. The response probabilities over 196 trials were taken as a "real" performance in the experiment. The process was repeated with 7 more prototype participants so that each sample of 16 contained 8 known prototype participants.

Similarly, each exemplar-based participant was given a random set of weights from among the 46,656 configurations in the exemplar model. These weights determined the

14 idealized categorization probabilities, which were then disturbed statistically by allowing chance to operate over 196 trials in the task. These probabilities were stored as a real participant's performance in the experiment, and the process was repeated with seven more known exemplar categorizers.

Finally, the eight known prototype performances and the eight known exemplar performances were aggregated, and both models were fit to each sample's aggregated performance. Seventy-four such samples of 16 participants were created and analyzed in this way. Many of these samples had characteristics (i.e., the fit of exemplar and prototype models to the performance of the prototype subgroup, the exemplar subgroup, and the aggregated data for the whole sample) that were like those we found in the real experiment.

The crucial result, though, is that the exemplar model fits the aggregated data better than the prototype model does, even when the sample includes eight known specimens of each strategy (fits of .0268 and .0478, respectively), $t(146) = 9.45, p < .05$.

It is a sobering thought that aggregating data favors the exemplar model, even when the participant population's provenance is known and known to be balanced. It raises the possibility that other aggregated analyses have found an advantage for the exemplar model by averaging away important pockets of prototype performance. Indeed, in some cases the reported advantage may have been more of an insight about the effects of aggregating over strategies than a comment on what strategies individual participants are or are not using. Nosofsky, Palmeri, and McKinley (1994) have also pointed out the risks inherent in fitting aggregated data and recommended finer scaled, individual-participant analyses (though not precisely the ones used here). In our data, individual-participant analyses yielded an equivalence between models.

### *Modeling Experiment 1's Other Conditions*

This equivalence is seen in all four conditions of Experiment 1. Over the whole group of 64 participants, 32 participants were fit best by each model. For the poorly differentiated NLS categories, moderately differentiated NLS categories, poorly differentiated LS categories, and moderately differentiated LS categories, respectively; there were 4, 8, 12, and 8 prototype participants. Across those same four conditions, the average fits for the prototype and exemplar models, respectively, were .3035 and .2513, .2048 and .2340, .2367 and .2726, and .1898 and .2042. One interesting possibility is that prototype strategies would be discouraged by the poorly structured categories. A hint of this is seen in the result that the poorly differentiated NLS condition contained the fewest prototype participants and the one overall fit disadvantage for the prototype model. However, the poorly differentiated LS condition clouds this issue because it produced the largest number of prototype participants and competitive fits by the prototype model.

The simplest conclusion is probably that all four conditions in Experiment 1 demonstrated an equivalence between the prototype and exemplar models, unlike the advantage for the exemplar model seen in previous research. Possibly, then, it is something that all four category structures share (a higher dimensionality of the stimulus space or more exemplars per category) that produces the stronger showing of the prototype model. In any case, this stronger showing, and the important subgroups of prototype participants, are the principal results of Experiment 1. In obtaining those results, the sensitivity of participant-by-participant models was critical.

## Experiment 2

Leaving Experiment 1, we faced two questions. First, it seemed the prototype model made a stronger showing here than in previous research, better capturing half of the performances. However, it was not possible to unpack the aggregated data in Medin's original research to make a direct comparison. Second, and worse, on aggregating the data, we showed the same advantage Medin had found for the exemplar model. This raised the possibility that previous studies had averaged away variations in participants' strategies.

In Experiment 2 we addressed both issues. To do so, we compared participants' performance on the NLS categories of Medin and Schwanenflugel (1981, Experiment 2) and the present moderately structured NLS categories. This allowed us to compare directly NLS category structures with fewer or more exemplars per category and with poorer or better category structure, with the same procedure, the same instructions, and the same duration of training. This also allowed individual-participant analyses that published reports do not support after the fact, but which do seem valuable.

One prediction would be that substantial numbers of participants, given participant-by-participant analyses, would turn out to adopt prototype strategies even given the original categories of Medin and Schwanenflugel (1981). This would confirm that aggregating performance can conceal an equivalence between models and would raise important issues about previous studies on NLS categories.

However, our favored prediction was that poorly differentiated category structures with few exemplars would undercut prototype strategies, encourage exemplar strategies, and produce fairly homogeneous performance with little prototype advantage and decent exception–item performance. Consequently, we predicted that most individual participants would be best fit by the exemplar model. This result would justify the claims from previous studies that were based on modeling aggregated data.

In contrast, we thought that the NLS condition of Experiment 1 (better differentiated and better stocked with exemplars) would encourage prototype strategies, and produce heterogeneous performance with strong prototype performance and poor exception–item performance. Consequently, we predicted that the data from many individual participants would be best fit by the prototype model, even if the aggregated data for the whole sample misleadingly favored the exemplar model.

## Method

*Participants.* Thirty-two introductory psychology students participated to partially fulfill a course requirement.

*Stimuli and category structures.* In this experiment we contrasted a four-dimensional stimulus set with poor category differentiation (like that used by Medin & Schwanenflugel, 1981, Experiment 2) and a six-dimensional stimulus set with better category structure (like that used in Experiment 1). Sixteen participants received each stimulus set.

The four-dimensional stimuli were derived from four prototype pairs (*buno–kypa, daki–sego, mufa–vosy,* and *leta–giru*). The first member of each pair was designated as stimulus 0 0 0 0, the second as stimulus 1 1 1 1. These prototypes were created subject to the constraints described in Experiment 1. The Category A members were 0 0 0 0, 0 0 0 1, 0 1 0 0, and 1 0 1 1. The Category B members were 1 1 1 1, 1 0 1 0, 0 1 1 1, and 1 0 0 0. This category structure, used by Medin and Schwanenflugel (1981, Experiment 2), included the complementary pairs of items (A3 and A4 and B3 and B4) and the exception items (A4 and B4) that are important features of these NLS categories. The six-dimensional stimuli were derived from four prototype pairs (*hafudo–nivety, gafuzi–wysero, banuly–kepiro,* and *lotina–gerupy*) that used the abstract structure shown in the Appendix.

*Procedure.* Participants were tested individually. Words were presented on a computer terminal in blocks of eight trials (four-dimensional) or 14 trials (six-dimensional)–each block a random permutation of all the stimuli in the experiment. Participants responded by using the 1 and the 2 keys on the number keypad. Correct responses were rewarded by a brief whooping sound generated by the computer; errors earned a 1-s low buzzing sound. A running total of participants' correct responses was displayed at the top of the screen. Trials continued in unbroken fashion until 392 trials had been presented (49 blocks for the four-dimensional stimuli and 28 blocks for the six-dimensional stimuli). Entering the experiment, participants were told that they would see nonsense word stimuli that could be classified as 'Group 1' words or 'Group 2' words. They were further told to

> Look carefully at each word and decide if it belongs to Group 1 or Group 2. Type a '1' on the keypad if you think it is a Group 1 word and a '2' if you think it is a Group 2 word. If you choose correctly, you will hear a 'whoop' sound. If you choose incorrectly, you will hear a low buzzing sound. At first, the task will seem quite difficult, but with time and practice, you should be able to answer correctly.

## Results

Performance on both category structures was reasonably good, with 73% correct overall for the six-dimensional categories (79% over the last 168 trials) and 66% correct overall for the four-dimensional categories (73% over the last 168 trials). Figure 6A shows correct percentages by 56-trial blocks over the course of the experiment.

However, as discussed in Experiment 1, overall performance levels blur different performance profiles and may blur different performance strategies by participants. Figures 6B and 6C show for both category structures the average percentage correct by trial block for prototypes and exceptions. For the Medin and Schwanenflugel (1981) categories, performance was fairly homogeneous over the range of typicalities presented in each category. This was definitely not the case for the six-dimensional categories. To confirm this effect, we entered the data into a three-way ANOVA with prototype–exception and trial block as within-subject variables and four dimensions–six dimensions as a between-subjects variable. There was a significant interaction in the ANOVA between the number of dimensions and item type, $F(1, 30) = 31.30$, $p < .05$, $MSE = 0.111$. From the four-dimensional to the six-dimensional stimulus sets participants became more accurate on the prototype items (71% correct vs. 86% correct) but less accurate on the exception items (56% correct vs. 36% correct).

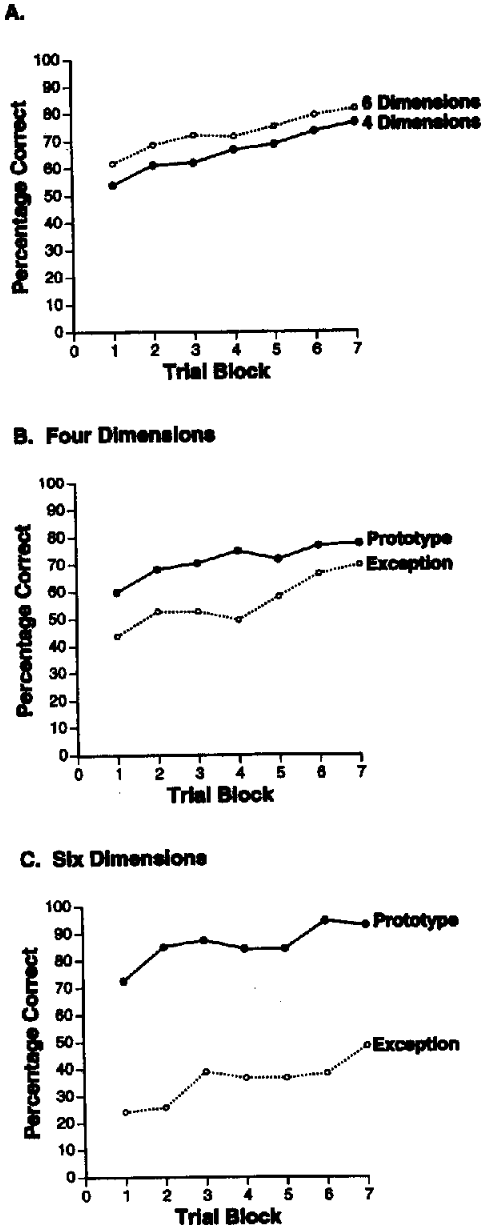

*Figure 6.* (A) Performance by 56-trial blocks on the six-dimensional and four-dimensional categories of Experiment 2. Performance on prototype and exception items by 56-trial blocks for the four-dimensional categories (B) and the six-dimensional categories (C).

As in Experiment 1, formal modeling evaluated whether the fit of different models might suggest that participants were relying on different strategies in the two tasks. The procedures for modeling were those described in Experiment 1, and they focused on participants' last 168 trials: 12 blocks of 14 stimuli for the six-dimensional categories and 21 blocks of 8 stimuli for the four-dimensional categories.

### *Modeling the Six-Dimensional Case*

*Fits to individual participants' data.* The mean fits averaged over participants were the same for the prototype and exemplar models (.2494 and .2263, respectively). Just as in Experiment 1, the prototype model and the exemplar model each fit better the performance of 8 participants.

*The prototype subgroup.* The 8 participants who were fit better by the prototype model averaged 93%, 80%, and 23% correct on the prototypes, normal items, and exceptions, respectively. On average, the prototype model fit these participants' data better than did the exemplar model (.1617 and .2980, respectively). Figure 7A shows the average of the eight observed performance profiles and the average of the eight best fitting profiles. Again, the prototype model tends to grant prototypes their observed advantage and simultaneously predicts bad exception–item performance. In contrast, Figure 7B shows that the exemplar model fits less well partly because it persistently underpredicts the prototypes' performance and overpredicts exception–item performance. The two models, respectively, heterogenize performance appropriately or homogenize it inappropriately.

These prototype participants are not just statistical fluctuations around the configuration of the exemplar model that fits best their combined performance. Illustrating this, we ran a simulation containing 500,000 exemplar-based categorizers. Each simulated categorizer was given the 14 categorization probabilities of the best fitting configuration and was

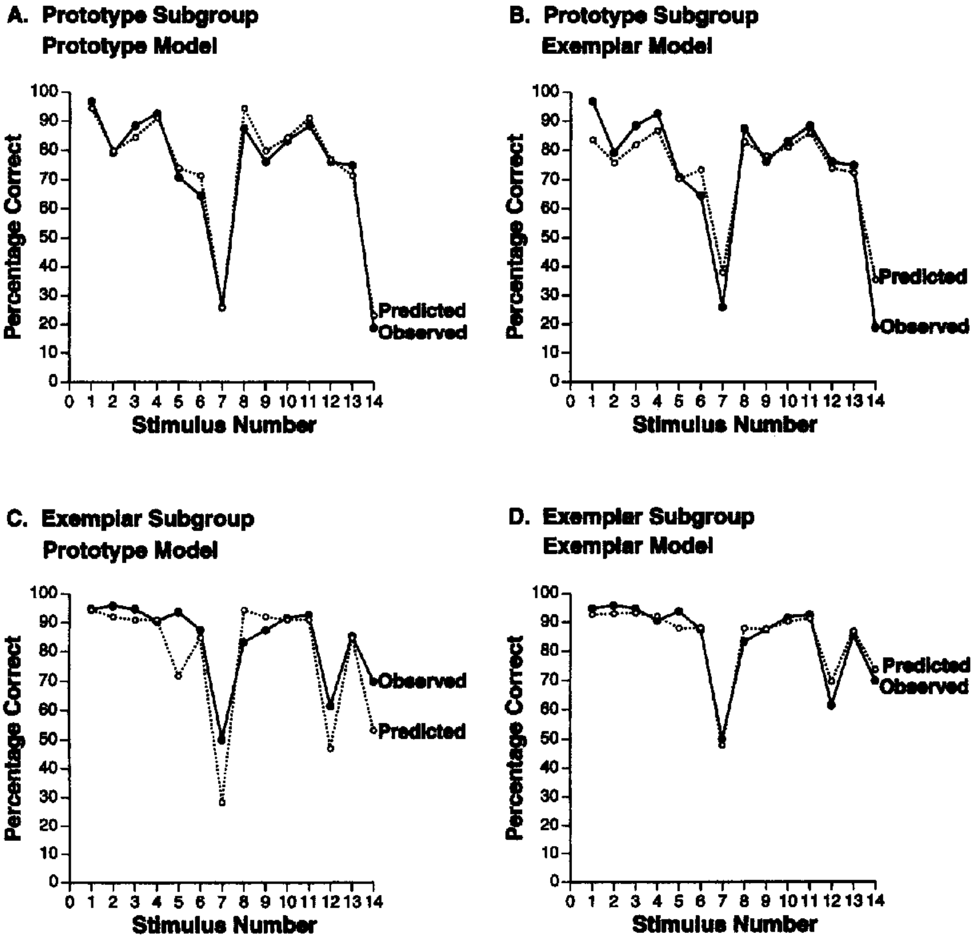

*Figure 7.* (A) The fit of the prototype model to the performances of the prototype participants for Experiment 2's six-dimensional categories. The solid line shows the average of the observed performance profiles produced by the 8 prototype participants. Each participant's performance was then modeled individually, and the dashed line shows the average of the best fitting predicted performance profiles. (B) The fit of the exemplar model to the performances of the prototype participants. (C) The fit of the prototype model to the performances of the exemplar participants for Experiment 2's six-dimensional categories. (D) The fit of the exemplar model to the performances of the exemplar participants.

allowed to perform 168 trials in our task (with chance operating). A simulated observer appeared only eight times in a thousand that performed as well on the prototypes as the subgroup but as poorly on the exceptions.

*The exemplar subgroup.* The 8 participants who were fit better by the exemplar model averaged 89%, 88%, and 60% on the prototypes, normal items, and exceptions, respectively. On average, the exemplar model fit these participants' data better than the prototype model did (.1547 and .3372, respectively). The prototype model fit these data poorly partly because it seriously underpredicted exception–item performance (Figure 7C). The exemplar model fared better (Figure 7D). The statistical procedures already described confirmed that these participants were unlikely to be just statistical fluctuations of the prototype model's best fitting configuration. Only three times in a hundred did that configuration produce, by chance, a pattern combining prototype performance as bad as these 8 participants achieved with exception performance as good as they achieved. Thus, both the individual-participant analyses and the subgroup analyses revealed the same parity between models as seen in Experiment 1.

*Fits to aggregated performance data.* Even so, we found that the exemplar model fit the aggregated data for all 16 participants better than did the prototype model (.0342 and .0500, respectively). This underscores the caution that modeling aggregated data can mislead by homogenizing away prototype processing.

### *Modeling the Four-Dimensional Case*

Thirteen of 16 participants were fit best by the exemplar model. On average, the exemplar model fit participants' data profoundly better than the prototype model did (.0872 and .4976, respectively). Figure 8A shows that the prototype model fails partly for severely underpredicting exception–item performance. (The prototypes are Stimuli 1 and 5 in the figures; the exception items are Stimuli 4 and 8.) The exemplar model captures these data far better (Figure 8B).

Across conditions in Experiment 2, a two-way ANOVA

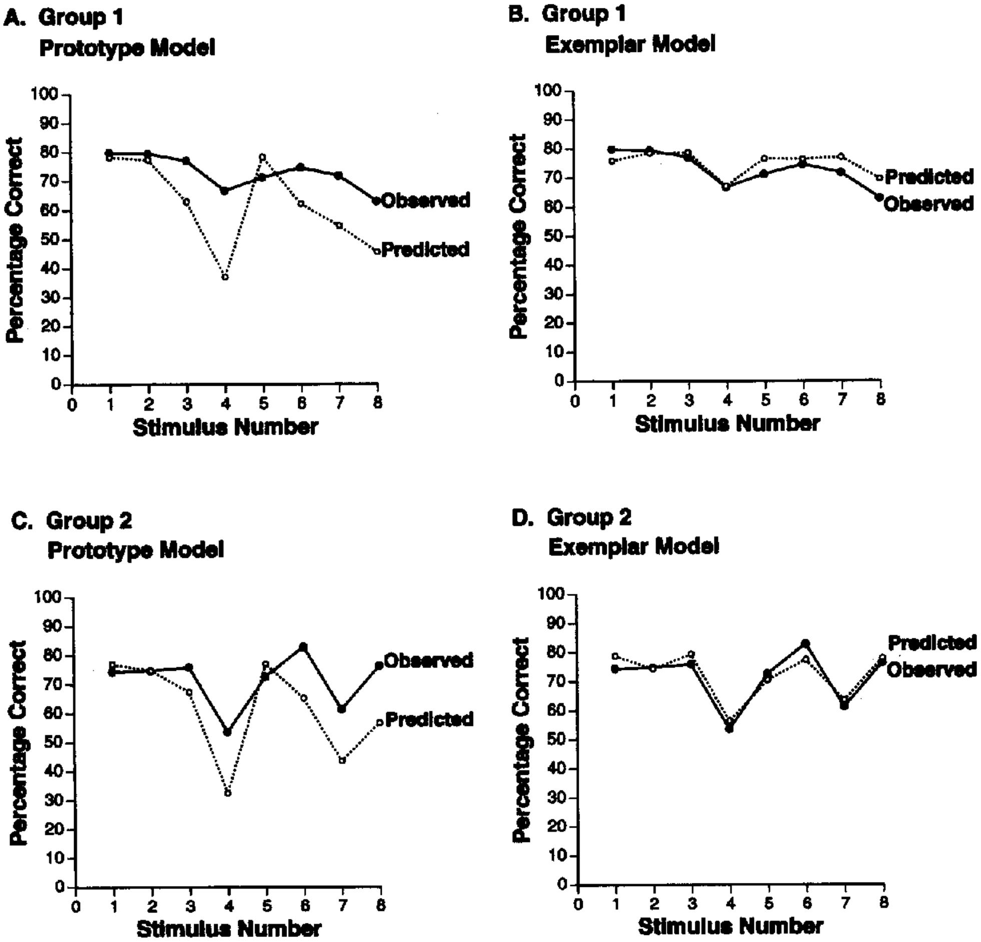

*Figure 8.* (A) The fit of the prototype model to the performances on Experiment 2's four-dimensional categories. The solid line shows the average of the 16 observed performance profiles. Each participant's performance was then modeled individually, and the dashed line shows the average of the best fitting predicted performance profiles. (B) The fit of the exemplar model to those same performances, graphed as in 8A. (C) The fit of the prototype model to the performances in the confirmatory sample of Experiment 2. (D) The fit of the exemplar model to those performances.

was conducted on the fits of the models by participant with category type (four dimensional vs. six dimensional) as a between-subjects variable and model type (exemplar vs. prototype) as a within-subject variable. This analysis revealed a significant interaction, $F(1, 30) = 8.79$, $p < .05$, $MSE = 0.068$, indicating that the exemplar model was strongly advantaged over the prototype model only in the case of the four-dimensional stimulus sets.

To emphasize the advantage of the exemplar model in the four-dimensional case, we tested two samples of 16 participants on those categories. In the second case also 13 of 16 participants were fit better by the exemplar model, and once again only the exemplar model fit the overall data pattern successfully (Figures 8C and 8D).

Thus, Experiment 2 provided several useful kinds of information. It confirmed the advantage of the exemplar model given the poorly differentiated, exemplar-poor categories of Medin and Schwanenflugel (1981). In fact, this is a far larger advantage for the exemplar model than Medin and his coworkers typically found. Medin and Schwanenflugel's original conclusions are safe from concerns that potentially attend the modeling of aggregated data, in the category structures they tested. Experiment 2 also confirmed the equivalence between the models given a better differentiated, more exemplar-rich category structure, and underscored the importance of participant-by-participant modeling analyses that appear to be more sensitive and appropriate.

## General Discussion

### *When Models Falter*

Figure 9 summarizes the present data and illustrates conditions that challenge the prototype and exemplar models. As exception–item performance improves, the fit of the prototype model is compromised (Figure 9A, $r = .88$),

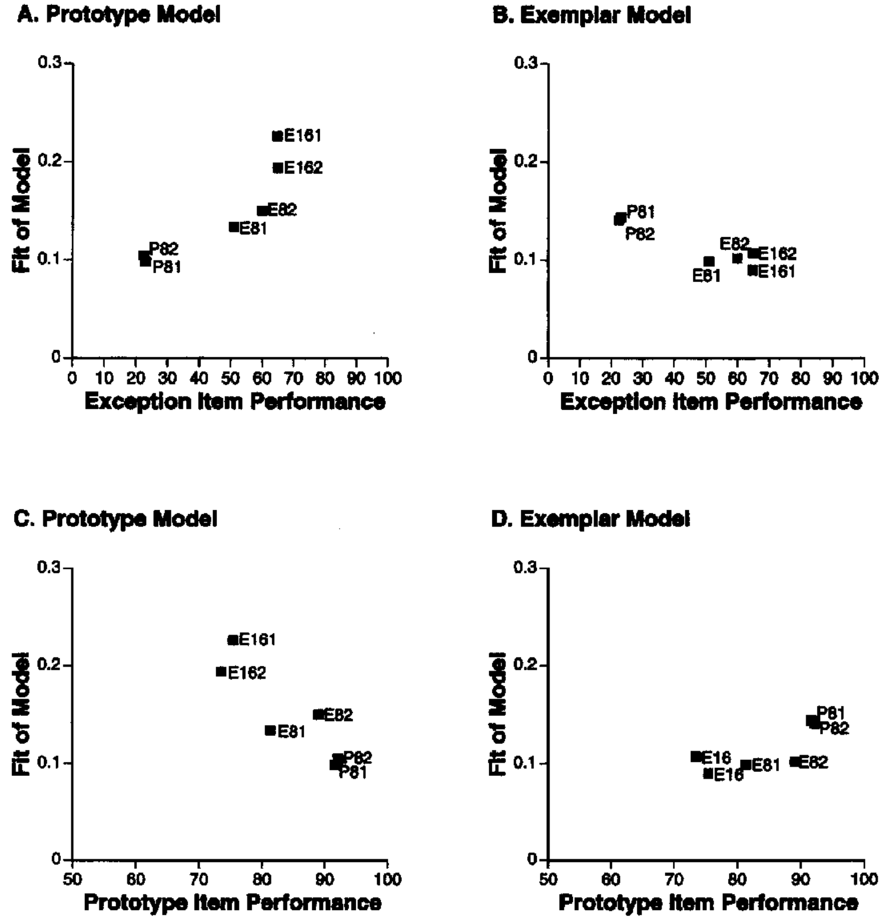

*Figure 9.* (A) The fit of the prototype model to performance profiles that show varying levels of exception–item performance. Shown are the prototype subgroups from Experiments 1 and 2 (P81 and P82), the exemplar subgroups from Experiments 1 and 2 (E81 and E82), and the whole groups that received the four-dimensional categories in Experiment 2 (E161 and E162). To blend more smoothly the fits to performances based on 8 and 14 exemplars, we used the root mean square deviation as a measure of fit. (B) The fit of the exemplar model to the performance of these same groups. (C) The fit of the prototype model to the same performance profiles, but now showing their varying levels of prototype performance. (D) The fit of the exemplar model to the performance of these same groups.

whereas the exemplar model fits better (Figure 9B, $r = -.94$). Conversely, as prototype performance improves, the prototype model fits better (Figure 9C, $r = -.88$), whereas the exemplar model falters (Figure 9D, $r = .75$).

Thus the exemplar model performs worse when fitting heterogenized performance, with larger prototype effects and lower exception–item performance. This is why it fits poorly the performance profiles of the prototype participants (Figures 5 and 7) and fits well the performances shown in Figure 8. The model has this character because it lets exemplars retrieve themselves, and this self-retrieval can support correct categorizations. The prototypes and the exception items, experienced repeatedly in training, can both rely on these self-retrieval episodes and this brings their levels of performance closer together. Moreover, even though the prototype is surrounded by similar exemplars, the one-dimensional mismatches can reduce multiplicative similarity to the point where "close" neighbors contribute negligibly to processing and hardly boost prototype performance. This is especially true for high sensitivity (low-similarity parameters) in the exemplar model and for the present case in which the prototypes were experienced in training.

In contrast, the prototype model performs better when prototype and exception–item performance diverge most. Calling an exceptional category member a nonmember, and making NLS categories into LS ones by reassigning exceptions, are the specialties of the prototype model. The model has this character because there is no exemplar self-retrieval that can increase exception performance. Also, the prototype has perfect self-similarity, producing strong prototype effects.

### *Interpreting the Performance of Prototype Participants*

Figure 10A shows the prototype and exception–item performance for each of the 16 prototype participants in Experiments 1 and 2. These participants have failed to learn the six-dimensional NLS categories, in the sense favored by Medin and his coworkers, of being able to accept all category members and reject all nonmembers (Medin & Schwanenflugel, 1981; Wattenmaker et al., 1986). Their performance shows the problematic nature of the exceptions and the reallocation of these items into the opposing categories. Through this reallocation, participants have turned the NLS categories into good LS ones. To do so, they have ignored errors, feedback, the experimenter's category definitions, and even the possibility of exemplar memorization. For them, there is a linear separability constraint at work. This is why the prototype model accounts easily, naturally, and best for these performances. Emphasizing this point, Figure 10B samples the entire constellation of performance profiles that are producible by the prototype model. This figure was created by moving systematically through all 1,115,730 parameter settings of a fine grained scan of the prototype model and capturing the average prototype performance and exception–item performance predicted by a random one in three hundred of these simulated observers. Prototype participants lie comfortably in the region of performance space covered by the prototype model.

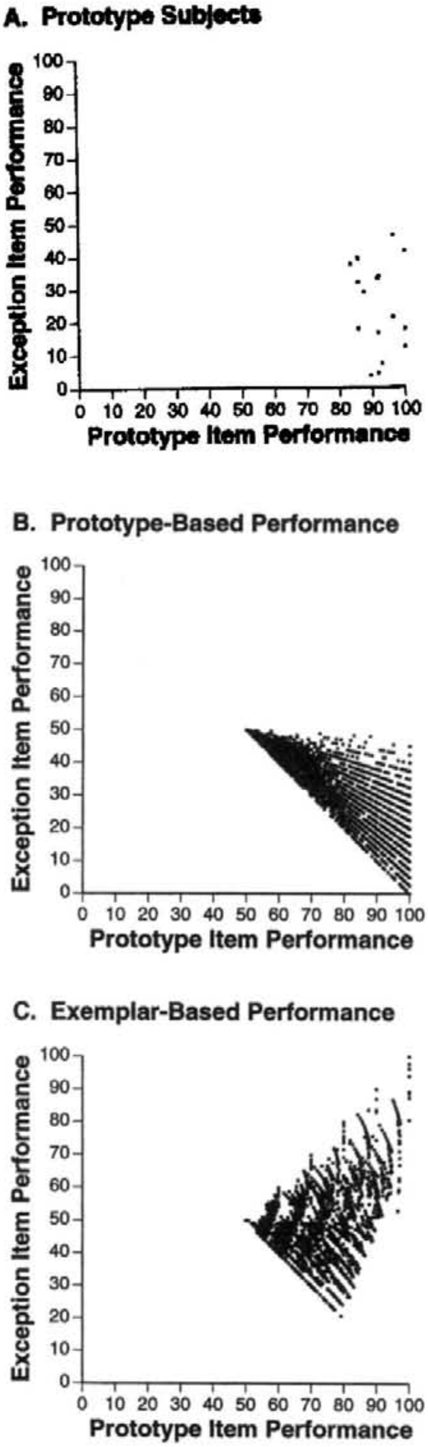

*Figure 10.* (A) Average prototype–item performance and exception–item performance for the 16 prototype participants from Experiments 1 and 2. (B) Average prototype–item performance and exception–item performance for about 3,700 parameter settings of the prototype model, randomly chosen during a detailed scan of its parameter space. (C) Average prototype–item performance and exception–item performance for about 5,400 parameter settings of the exemplar model, randomly chosen during a detailed scan of its parameter space.

In contrast, the performance patterns of the prototype participants fit poorly with some ideas about exemplar-based categorization strategies. One wonders why the participants did not just memorize the exemplars, especially the exceptions, and let these traces guide the correct categorization. One wonders why they did not code the problematic stimuli more relationally, noting the appropriate label for that particular constellation of features. Even after nearly 400 trials (in Experiment 2), they have not done so.

Emphasizing this point, Figure 10C samples the entire constellation of performance profiles that are producible by the exemplar model. This figure was created by moving systematically through all 19,487,171 parameter settings of a fine grained scan of the exemplar model and capturing the average prototype performance and exception–item performance predicted by a random 1 in 5,000 of these simulated observers. For this purpose, to ensure that no relevant performance profiles were excluded, the exemplar model was granted a seventh free parameter that evaluated different rates of guessing. As shown by the figure, many prototype participants lie outside this region of performance space (and the earlier simulations confirmed that chance events do not explain this). The exemplar model can predict exception–item performance as low as that shown by some of these participants. However, then it seriously underpredicts their prototype performance. The exemplar model can predict prototype performance as high as that shown by these participants. However, then it seriously overpredicts their exception–item performance. The prototype participants were fit poorly because they required the model to accomplish two goals that are incompatible given its behavior. The simplest interpretation is that prototype participants were using a prototype strategy that subjected them to a linear separability constraint.

Illustrating the existence of these participants, and the linear separability constraint they operate under, was the primary goal of the present research, for it represents a useful counterweight to existing research. The present category structures have not induced strong prototype strategies and an LS constraint in all participants, but clearly that approach gains (and reaches parity) in the transition from the category structures of previous research (e.g., the four-dimensional categories of Experiment 2) to the six-dimensional categories of Experiments 1 and 2. This parity is not just a technical or formal equivalence born, for example, from the ceiling performance that both models would predict at high structural ratios. Rather, the exemplar model is failing in a predictable way to capture the performance profiles of many participants. These prototype participants are important because they have been an endangered species lately in the categorization literature.

### *Different Category Structures That Foster Different Task Psychologies*

Why is the prototype strategy more prominent in our data than in previous research? Structural differences between stimulus sets must underlie these differences. The four-dimensional stimulus sets used here and in many previous studies have fewer stimuli that are repeated sooner and more often during the experiment. They contain less differentiated categories and less deviant exception items. The structural ratio of 1.32 means that similarity is fairly homogeneously distributed between and within categories. The six-dimensional stimulus set used here had the opposite character in every respect.

These structural differences could produce different task psychologies or strategies. With fewer stimuli, exemplar memorization will be easier. With more repetitions, that strategy will be more obvious, even as the specific traces are reinforced. With less differentiated categories, the urge to form LS exemplar clusters will be weakened. For example, no big dissimilarity jumps between successive stimuli will signal that there are two well-organized groups of exemplars. With less deviant exceptions, there will be less "remorse" about including them in the correct category. The present six-dimensional stimulus set, by contrast, featured more stimuli to bookkeep, fewer repetitions to aid this process, more pull to the prototypes, and more problematic exceptions.

Further research will be needed to pinpoint which of these structural differences produced the strategy differences observed across conditions. For example, one might be able to manipulate independently the dimensionality of the stimulus set, the number of exemplars in categories, and the similarity structure of the categories, to see which factor is implicated most strongly. The balance between prototype and exemplar models, in all four conditions of Experiment 1, could suggest that the larger exemplar pools there either encouraged prototype strategies or discouraged exemplar strategies. However, possibly against this idea, McKinley and Nosofsky (1995) found that participants, after thousands of training trials, were able to defend nonlinear decision boundaries even given scores of unique stimuli.

Another possibility is that assimilation to prototypes grows stronger at better levels of category structure so that item-specific information is lost that could support the correct categorization of exceptions. However, against this idea is the fact that the modest structural ratio manipulation in Experiment 1 did not produce this effect. More generally, we believe that error signals are a powerful signal to participants that they need to adopt new strategies in categorization.

This discussion leads us to contrast two different configurations of categorization research. One configuration features small exemplar pools, poorly differentiated categories, learning difficulties, weak prototype effects, relatively strong exception–item performance, homogenized performance overall, and the comfortable fit of the exemplar model, plausibly indicating exemplar-based strategies. Such strategies would have a sound information-processing basis. As prototype strategies are undermined by impoverished similarity relationships, frequent error signals could prompt strategy changes, while frequent exemplar repetition reminds the participant about exemplar-memorization strategies and even builds the strongly individuated traces that serve them. Thus, this configuration of category structure, information-processing strategy, performance profile, and best fitting model might be a mutually reinforcing, closed system.

Another configuration of categorization research involves larger exemplar pools, well-differentiated categories, easier learning, strong prototype effects, poor exception–item performance, heterogenized performance overall, and the comfortable fit of the prototype model, plausibly indicating prototype-based strategies. Again, these strategies would have a sound information-processing basis. As prototype strategies succeed, rare error signals would discourage strategy changes, while larger exemplar pools camouflage and undermine the exemplar-based alternatives. This configuration of structure, processing, performance, and model could also be a mutually reinforcing, closed system. We do not prejudge which configuration is more generally applicable and true—either in the laboratory or in the natural ecology. Our point is only that both configurations have potential importance, both in the laboratory and in the natural ecology.

However, one must explore fully the broad range of category structures to know if this is true. It continues to be fashionable to inflict extraordinarily difficult category structures on participants. These structures do allow beautiful stimulus control and elegant modeling, but there may be a cost. Just as objects look red in red light, so category learning may look exemplar based for as long as only an exemplar-sparse, poorly differentiated light is shed on it.

### *Epochs of Learning and Models of Categorization*

Another perspective on our results comes from supposing that we observed participants during some early phase of learning (lasting through 392 trials), when the prototype participants were only about 25% correct on the exceptions. Even though we gave participants more trials than in the directly comparable studies, related experiments have sometimes given thousands of trials. If learning had gone on much longer, would not participants have mastered the exceptions finally, and would not the exemplar model have fit better then?

This possibility raises important issues. If one made participants persist until they learned the exceptions, the homogeneous performance landscape that resulted would favor the exemplar model. This advantage is not just circular—even though the criterion for ceasing training would be a comfortable configuration for the exemplar model (i.e., one in the top-right quadrant of the graphs in Figure 10). Participants' success would show that they can eventually transcend the difficulty of the exception items, possibly by applying special learning algorithms (e.g., exemplar memorization, etc.). However, equally interesting is the fact that here many participants made little headway with the exceptions even after 28 presentations of them. These participants are in a serious rut with prototypy and linear separability, and this is an extended stage of learning, even if they eventually transcend it.

The implication of this is that different training regimens used in category research make different commitments. For example, Nosofsky (1986), hoping to model a fully mature performance, tested participants on 1,200-trial blocks until they performed better than chance on all stimuli in the last 600 trials of a block. (In other studies Nosofsky modeled performance earlier in learning; e.g., see Nosofsky, 1989.) This extensive training eliminates poor performance on any "exception" items, homogenizes performance in the manner relevant here, and favors the exemplar model appropriately because of participants' eventual success. However, this training regimen chooses not to ask if participants went through important prototype-based epochs earlier in training.

If one chooses to ask, one may find different kinds of processes, and different kinds of best fitting models, during different epochs of training and learning. Studying eventual transcendance over exceptions is not the only important issue in categorization research. Default assumptions, first principles, general preferences, and mode chosen given cooperative circumstances—these are all equally plausible areas of investigation. Curiously, all of these could be prototype based, and all of these could enforce a linear separability constraint, but it could still be true that intense training would transcend the limits of these strategies and constraints and bring favor on the exemplar model. These default or preferred strategies have received scant research attention—at least until very recently.

For example, Ahn and Medin (1992) and Nosofsky et al. (1994) both suggested that participants learn categories by first making a straight, linear cut through the stimulus world, using a single-dimensional rule (see also results from sorting paradigms in Medin, Wattenmaker, & Hampson, 1987; Regehr & Brooks, 1995). Then, to the extent that the rule produces errors, participants invoke special learning algorithms to master the exceptions. These algorithms can involve exemplar memorization or the relational coding of featural complexes that need special classificatory processing. In an elegant demonstration of different learning epochs, Nosofsky et al. considered category structures that allowed generalization by either independent informative cues or by relational cues (correlated attributes). They showed that generalizations made early in learning followed individual diagnostic cues, whereas later generalizations followed relational coding.

These descriptions of categorization deemphasize the static context model and the idea that humans are pure, relational-coding creatures with exemplar-blotter memories. They emphasize that different epochs of learning involve different strategies and varieties of stimulus coding. They favor the idea that the default options and early epochs involve a strong linear separability assumption. They link relational coding to the special learning algorithms that come later in learning, and only if necessary. They suggest that the exemplar model will find progressively more favor later in learning because its kind of processing is allied to the eventual-transcendance strategies that conquer exceptions.

We endorse each of these conclusions. In turn, these models endorse our view about the problems and the processing consequences of poor category differentiation. Poorer category differentiation will guarantee less adequate rules, demand more focus on their exceptions, require more relational coding, and may well favor the exemplar model more. The only suggestion we would add is that early rules need not just be unidimensional. We believe that participants can also rapidly appreciate multidimensional hypotheses or rules and then set to work on the exceptions to these prototypes. This phenomenon may be best seen by providing better differentiated category structures—we believe that prototype strategies may especially accrete around well-differentiated prototype cores.

These converging perspectives on categorization are welcome. At times the combination of poorly differentiated category structures and demonstrations of eventual transcendance may have confined theories of categorization too much. Researchers have often placed participants in a difficult, exemplar-sparse corner of the domain of categorization, where exemplar models may be favored, and where prototype strategies may not even work. But then, researchers have patiently kept training while participants master these difficult categories. They do, and exemplar models fit well then because they should. This result is an important existence proof about participants' ultimate capacities in categorization. However, it may not be a subsistence proof about what humans do during categorization's "business as usual" in the natural ecology. Subsistence categorization, not just tour de force categorization, deserves researchers' careful study too.

### The Real LS Advantage

This distinction leads us to move beyond our particular results and to close by reconsidering some prevalent ideas in the literature today—that there is a kind of equivalence between LS and NLS categories, that humans do not benefit from linear separability in learning categories, and that there is no linear separability advantage or constraint (Ashby & Gott, 1988, p. 51; Kemler Nelson, 1984, p. 747; Lingle et al., 1984, pp. 93–94; McKinley & Nosofsky, 1995; Medin & Schwanenflugel, 1981; Murphy & Medin, 1985, p. 295; Wattenmaker et al., 1986, pp. 160–162).

These ideas have many roots. Demonstrating eventual transcendence highlights equivalence because participants eventually do learn NLS categories. It is easy to forget the extensive training that attends their learning. The original LS–NLS comparisons highlighted equivalence because NLS

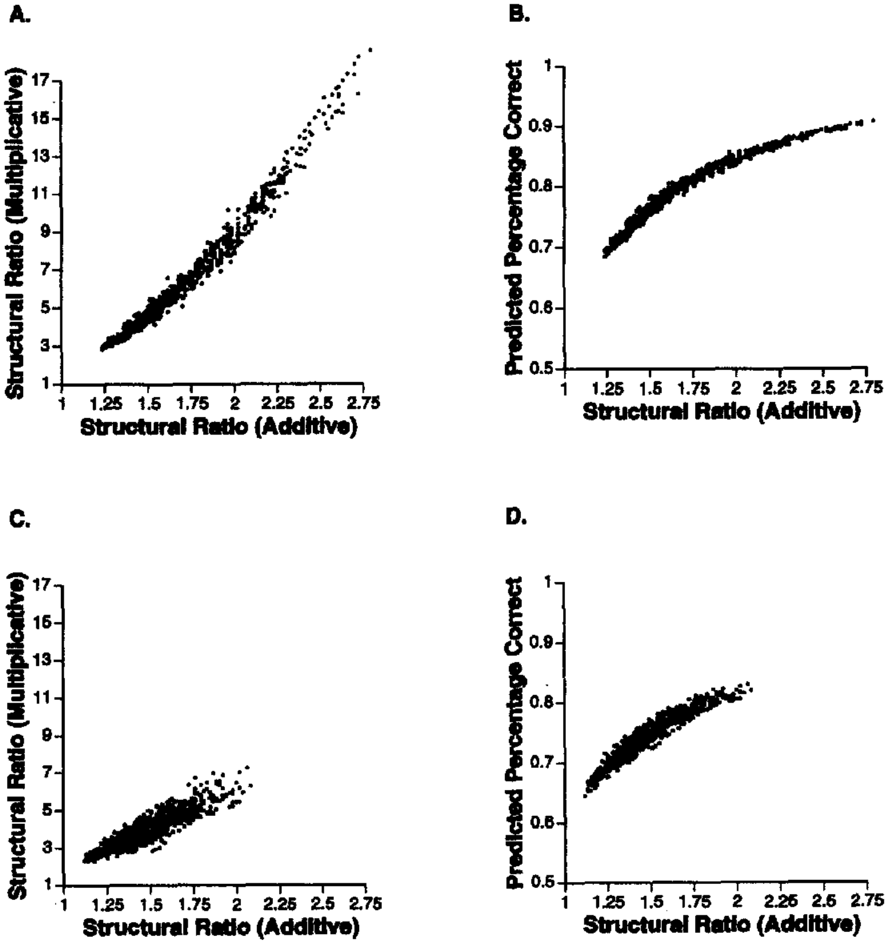

*Figure 11.* (A) Seven hundred hypothetical category structures illustrating the range of category differentiation offered by linearly separable categories. (B) The overall performance of a generic configuration of the exemplar context model (dimensional weights = .45) on those 700 LS stimulus sets. (C) Seven hundred hypothetical category structures illustrating the range of category differentiation offered by nonlinearly separable categories. (D) The overall performance of the same configuration of the exemplar model on those 700 NLS stimulus sets.

categories did not produce poorer learning there. It is easy to forget the skillful balancing of similarity relations that made equivalent performance the natural outcome of these experiments, and it is easy to drift beyond the careful specificity of the original claims in this area. Following on the research, equivalence was suggested because exemplar memorization could balance prototypy and support the learning of NLS categories, because relational-coding schemes could let participants encode strongly individuated exemplars and categorize them correctly, because the correlations between attributes could support the mastery of some NLS categories, and because conceptual and theoretical knowledge could also support the learning of NLS categories that are poorly structured perceptually.

In contrast, Figure 11 illustrates a basic nonequivalence between LS and NLS category structures. Figure 11A shows 700 hypothetical linearly separable category structures by structural ratio, and Figure 11B shows how well a generic exemplar-based categorizer (with parameter settings of .45 on each dimension) would perform overall on each of them. Not surprisingly, there is a strong relationship between overall performance and category differentiation. The same would be true of a prototype-based categorizer. The morel is that mushroom foragers should fervently hope that edible and poisonous mushroom categories will be well differentiated. Predator avoiders should fervently hope that eagle and vulture categories will be well differentiated. If they are, the species will eat better—longer.

Figure 11C shows the similarity relationships for 700 hypothetical NLS category structures constructed as here, with one exception item in each category. Figure 11D shows the performance of the exemplar-based categorizer again. Figure 11 makes a point that is sometimes lost in studies of categorization and linear separability. On average, LS categories will be better differentiated. On average, LS categories will be more easily learned and performed at a higher level. The fact of NLS weakens the similarity relationships within categories and reduces the structural ratio for a pair of categories. The fact of NLS damages the performance of the exemplar-based (and prototype-based) categorizer. The 75% levels of performance shown in Figure 11D could spell real disaster for a real species that mistook, one fourth of the time, poisonous mushrooms for edible ones and eagles for vultures. The species would eat worse—and worse, be eaten.

Furthermore, in a case like this the eventual capacity to transcend the poor perceptual structure and master the whole category is not the vital issue. The equality of performance on two carefully selected category structures that are perfectly balanced for similarity is also not the vital issue. Rather, the vital issue is how long do categories take to learn, and how difficult are they to learn, with these questions given urgency by the seriousness of mistakes in the natural world.

On this vital issue both exemplar-based and prototype-based models agree. Over the whole range of category structures available, as assessed by the two most prominent models of cognitive processing in categorization tasks, there has always been, and there will always be, a profound linear separability advantage.

## References

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

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 33–53.

Hartley, J., & Homa, D. (1981). Abstraction of stylistic concepts. *Journal of Experimental Psychology: Human Learning and Memory, 1,* 33–46.

Homa, D., Rhoads, D., & Chambliss, D. (1979). Evolution of conceptual structure. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 11–23.

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.

Kemler Nelson, D. G. (1984). The effect of intention on what concepts are acquired. *Journal of Verbal Learning and Verbal Behavior, 23,* 734–759.

Lingle, J. H., Altom, M. W., & Medin, D. L. (1984). Of cabbages and kings: Assessing the extendibility of natural object concept models to social things. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 71–117). Hillsdale, NJ: Erlbaum.

Locke, J. (1959). *An essay concerning human understanding* (A. C. Fraser, Ed.). New York: Dover. (Original work published 1690)

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior, 23,* 250–269.

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 128–148.

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 37–50.

Medin, D. L., Dewey, G. I., & Murphy, T. D. (1983). Relationships between item and category learning: Evidence that categorization is not automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 607–625.

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

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 75,* 355–368.

Medin, D. L., & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 241–253.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19,* 242–279.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology, 32,* 89–115.

Murphy, G. L., & Medin, D. L. (1985). Role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 104–114.

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

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of*

*Experimental Psychology: Learning, Memory, and Cognition, 13,* 87–108.

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception & Psychophysics, 45,* 279–290.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of William K. Estes* (pp. 149–167). Hillsdale, NJ: Erlbaum.

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

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 353–363.

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology, 83,* 304–308.

Regehr, G., & Brooks, L. (1995). Category organization in free classification: The organizing effect of an array of stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 347–363.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 111–144). New York: Academic Press.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology, 7,* 192–238.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7,* 573–605.

Sebestyen, G. S. (1962). *Decision-making processes in pattern recognition.* New York: Macmillan.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75* (13, Whole No. 517).

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decision. *Psychological Review, 81,* 214–241.

Smith, J. D. (1989). Analytic and holistic processes in categorization. In B. E. Shepp & S. Ballesteros (Eds.), *Object perception: Structure and processes* (pp. 297–323). Hillsdale, NJ: Erlbaum.

Smith, J. D., & Shapiro, J. H. (1989). The occurrence of holistic categorization. *Journal of Memory and Language, 28,* 386–399.

Smith, J. D., Tracy, J., & Murray, M. J. (1993). Depression and category learning. *Journal of Experimental Psychology: General, 122,* 331–346.

Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. L. (1986). Linear separability and concept learning: Context, relational properties, and concept naturalness. *Cognitive Psychology, 18,* 158–194.

Whittlesea, B. W. A. (1987). Preservation of specific experiences in the representation of general knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 3–17.

Whittlesea, B. W. A., Brooks, L., & Westcott, C. (1994). After the learning is over: Factors controlling the selective application of general and particular knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 259–274.

## Appendix

### Category Structures Used in Experiment 1

| Categories linearly separable | | | | Categories not linearly separable | | | |
|---|---|---|---|---|---|---|---|
| Category A | | Category B | | Category A | | Category B | |
| Structure | Stimuli | Structure | Stimuli | Structure | Stimuli | Structure | Stimuli |
| Difficult | | | | Difficult | | | |
| 101000 | LITUPO | 001111 | GITANY | 010000 | GERUPO | 011111 | GETANY |
| 100010 | LIRUNO | 011011 | GETUNY | 001000 | GITUPO | 101111 | LITANY |
| 100001 | LIRUPY | 011101 | GETAPY | 000001 | GIRUPY | 111101 | LETAPY |
| 011000 | GETUPO | 101011 | LITUNY | 101000 | LITUPO | 011110 | GETANO |
| 010100 | GERAPO | 101110 | LITANO | 100010 | LIRUNO | 101101 | LITAPY |
| 001001 | GITUPY | 110101 | LERAPY | 001001 | GITUPY | 111001 | LETUPY |
| 000011 | GIRUNY | 111100 | LETAPO | 110111 | LERANY | 000010 | GIRUNO |
| Easy | | | | Easy | | | |
| 000000 | BUNOLI | 111111 | KYPERA | 000000 | BUNOLI | 111111 | KYPERA |
| 100000 | KUNOLI | 110111 | KYNERA | 100000 | KUNOLI | 011111 | BYPERA |
| 010000 | BYNOLI | 111101 | KYPELA | 010000 | BYNOLI | 101111 | KUPERA |
| 100001 | KUNOLA | 010111 | BYNERA | 001000 | BUPOLI | 110111 | KYNERA |
| 011000 | BYPOLI | 011110 | BYPERI | 000010 | BUNORI | 111011 | KYPORA |
| 001010 | BUPORI | 101011 | KUPORA | 000001 | BUNOLA | 111110 | KYPERI |
| 000101 | BUNELA | 101110 | KUPERI | 111101 | KYPELA | 000100 | BUNELI |

Received September 14, 1995
Revision received August 7, 1996
Accepted August 15, 1996 ■

View publication stats