Memory & Cognition
1997, 25 (5), 625-640

Associative and similarity-based processes in categorization decisions

JAMES A. HAMPTON
City University, London, England

Two experiments were directed at distinguishing associative and similarity-based accounts of systematic differences in categorization time for different items in natural categories. Experiment 1 investigated the correlation of categorization time with three measures of instance centrality in a category. Production frequency (PF), rated typicality, and familiarity from category norms for British participants (Hampton & Gardiner, 1983) were used to predict mean categorization times for 531 words in 12 semantic categories. PF and typicality (but not familiarity) were found to make significant and independent contributions to categorization time. Error rates were related only to typicality (apart from errors made to ambiguous or unknown items). Experiment 2 provided a further dissociation of PF and typicality. Manipulating the difficulty of the task through the relatedness of the false items interacted primarily with the effect of typicality on categorization time, whereas, under conditions of easy discrimination, prior exposure to the category exemplars affected only the contribution of PF to the decision time. The dissociation of typicality and PF measures is interpreted as providing evidence that speeded categorization involves both retrieval of associations indexed by PF and a similarity-based decision process indexed by typicality.

The phenomenon of gradedness within categories is the finding that some instances of common taxonomic categories (e.g., robin as a bird) are consistently judged as more typical or representative of their categories than are others (Barsalou, 1985; Hampton, 1979; Rosch, 1975). Typical instances have been shown to receive preferential processing in a wide range of cognitive tasks (for a review, see Hampton, 1993). For example, in a speeded categorization task, typical words are categorized more rapidly and more accurately than are atypical words (Hampton, 1979; Smith, Shoben, & Rips, 1974). Typical instances also tend to be generated with a high production frequency when people are asked to retrieve examples of categories from memory (Battig & Montague, 1969; Hampton & Gardiner, 1983; Mervis, Catlin, & Rosch, 1976).

There are two fundamentally different ways of interpreting gradedness effects in common taxonomic categories. One is in terms of the learning history of the individual, and it proposes that "good" category members are those that have been most often associated with the category in the past. For example, nonanalytic models of category learning and concept representation (e.g., Brooks, 1978, 1987) would emphasize the importance of past associations in determining speed of categorization. The other interpretation is in terms of what Rosch (1975) called the *internal structure* of the category concept, according to which the "good" category members are those that share the greatest similarity with the prototypical representation of the category concept.

These two different ways of accounting for the gradedness of categories have been applied to explaining within-category variation in categorization times in models of semantic memory, which have taken this variation as reflecting one of two processes: search processes in an associative net, or decision processes involving comparison of the item with the category concept (see Chang, 1986, and Smith, 1978, for reviews). Search models, such as Glass and Holyoak's (1975) marker search model, relate within-category variation to frequency of co-occurrence, within a traditional associationist framework. Categorization depends on retrieving the correct relation from a network of prestored semantic relations, including both property statements such as "has legs" and category statements such as "is a bird." Frequency of use of a semantic relation determines its ease of retrieval, because frequently used links develop greater associative strength (Thorndike's "law of practice"). The alternative class of model, similarity-based comparison models of categorization such as Smith et al.'s (1974) characteristic feature model and McCloskey and Glucksberg's (1979) property comparison model, propose that, in a categorization decision task, the feature overlap between an instance and a category is computed. In the property comparison model, property overlap between instance and category is sampled until a sufficient weight of evidence has accumulated either for or against categorization. Highly typical instances are categorized more rapidly than are atypical instances, since evidence for a positive decision accrues more rapidly for these items. According to a

The author thanks Larry Barsalou, John Gardiner, Margaret Gardiner, David Green, Frank Keil, Robert Lorch, Michael McCloskey, and anonymous reviewers for comments and suggestions on the research, and Andy Ojukwu and Bill Fitzpatrick for assistance in data collection. Correspondence should be addressed to the author at the Psychology Department, City University, Northampton Square, London EC1V 0HB, England (j.a.hampton@city.ac.uk).

"pure" similarity comparison model, property information (e.g., "has legs") is stored with each concept, but category membership (e.g., "is a bird") is computed each time through computation of the degree of property match.

In sum, network search models attribute within-category variation in categorization time to variation in the strength of the associative "is a" link between the instance and the category, whereas similarity-based decision models attribute the variation to differences in the similarity between instance and category in terms of the overlap of their semantic features.[1] Rips, Smith, and Shoben (1975) and Smith (1978) argued that these two different modes of explanation reflect important theoretical differences in assumptions about memory structure—in particular, concerning the role played by frequency of association as opposed to semantic content in determining the operating characteristics of semantic memory.

This theoretical distinction is an instance of a more general distinction with many parallels in cognitive science. One possible cognitive architecture is associative, and it has operating dynamics driven primarily by processing experience and, in particular, by the laws of associative learning. The other type of architecture is one whose dynamics are primarily driven by content—specifically, the logical structure of the information contained within it. This general distinction emerges in a number of fields. For example, in syntactic processing, the logico-semantic approach of Pinker and Prince (1988) and Fodor and Pylyshyn (1988) contrasts with the parallel distributed processing models of Rumelhart and McClelland (1986) and Smolensky (1987, 1988). In theories of category learning, rational analysis of the internal structure of concept categories (Anderson, 1990, 1991) can be contrasted with pure learning models that store exemplar–category associations (Brooks, 1978, 1987; Medin & Schaffer, 1978; Nosofsky, 1988). Models of lexical memory have similarly been concerned with the issue of whether association strength or semantic relatedness are chiefly responsible for semantic priming effects (Shelton & Martin, 1992).

The category verification task is a task with the potential to provide direct evidence of the validity of these two general views of cognition. For example, recent papers by Chumbley (1986), Casey (1992), and Larochelle and Pineau (1994) have used evidence concerning the relative influence of associative versus content factors on categorization time to draw conclusions about the structure of semantic memory. The approach adopted to differentiate the roles of associative versus structural effects in semantic memory by these researchers, which will also be adopted here, has been to assume that category production frequency (PF) and typicality, although often strongly correlated within a category (Barsalou, 1985; Hampton, 1979; Hampton & Gardiner, 1983; Mervis et al., 1976), in fact reflect theoretically and empirically distinct aspects of category structure. It has been established that the two measures reflect statistically independent sources of variance (Hampton & Gardiner, 1983). Given that they reflect different aspects of category structure, the research has taken PF to be a relatively direct measure of the association strength of the instance–category relation, reflecting the accessibility of instance–category associative links and, hence, the ease with which the items can be retrieved as members. PF reflects the associative aspect of category structure that (following traditional associationist theories) would correspond to frequency of co-occurrence in the history of learning the category. Typicality, on the other hand, is taken as a measure of the conceptual similarity of a category member to the category prototype. It reflects the structure of the learned information, rather than the frequency of encountering it. The intercorrelation of the two measures results from the fact that typical category members also tend to be those that are most commonly encountered and, hence, are most readily accessible. The measures remain distinct because there may still be some members that are commonly encountered but are dissimilar from the prototype or, alternatively, others that are rarely encountered but are very similar to the prototype.

Typicality and PF have both been shown to correlate with differences in the speed of category decisions. Chumbley (1986), Conrad (1972), Hampton (1984), Loftus (1973), and Wilkins (1970), among others, have shown that instances with a higher PF are more rapidly categorized. This result has been generalized to false category statements by Glass and Holyoak (1975), using a modified generation task, in which subjects produced false completions to category sentences. Alternatively, Casey (1992), Hampton (1979), Larochelle and Pineau (1994), and Smith et al. (1974) found similar effects on categorization time for high- versus low-*typicality* instances. The more typical an instance is in a category, then the faster is a positive categorization; for false category statements, the more similar a nonmember is to a category, then the slower people are to reject it as a category member (Hampton, 1979).

While neither measure may be a "pure" index of the theoretical dimension that it is assumed to reflect (there is after all no adequate model of the category instance retrieval task, or of the typicality rating task), it may reasonably be assumed that since the production task clearly involves a search and retrieval process, and the rating of typicality involves a careful consideration of the degree of similarity between an instance and the rest of the category, the two measures should at least contain independent variance corresponding closely to these aspects of semantic memory structure.

On the basis of these assumptions, recent studies (Casey, 1992; Chumbley, 1986; Larochelle & Pineau, 1994) have used regression methods to investigate which dimensions of category structure best predict categorization time. Results from these studies, however, have not been consistent. Two ways of measuring categorization time have typically been used: one in which the instance is presented first, followed after a delay by a category name, and the other in which the category name is presented first, followed after a delay by a possible instance. Chumbley (1986) measured categorization time in both

orders and found that typicality as a variable had no unique predictive power in either condition. The best predictors of positive categorization times were measures related to category and instance dominance—the strength of associations from the instance to the category, or vice versa. Chumbley concluded that any effects of semantic content (i.e., typicality) were therefore mediated through the associational structure of semantic memory, built up through co-occurrence of instances with their associated categories. However, a partial replication by Casey (1992) failed to find the same results. In Casey's study, typicality was a significant predictor variable in all experimental conditions, whereas category dominance and instance dominance were largely predictive only in the corresponding order conditions (see also Loftus & Scheff, 1971) in which they would predict the likelihood of successfully guessing the second word. In a third study that attempted to resolve this inconsistency, Larochelle and Pineau (1994) found results that largely replicated those of Casey. Typicality was the strongest predictor of categorization times, whereas category dominance again played a role only in the instance–category presentation order, where subjects would have been more able to guess the true category before it appeared when the instance–category pair had high category dominance. Larochelle and Pineau carefully review methodological differences among the different studies that could explain the discrepancy in results. Among these differences, key points appear to be methods for selection of materials, the validity of the normative measures used, and priming effects arising from the repetition of the same items within the experiment. (Clearly, if the same decision is made repeatedly, later decisions may be made by retrieving the earlier result, rather than running the decision or retrieval process de novo.) The inconsistent results in the literature point up the need for particular care when using regression methods. The selection of instances in each category needs to be representative to allow each variable its natural range of variation. The three studies cited used between four and eight instances per category, which is an insufficient sample size to properly represent the distribution of the independent variables within each category. Within-category variation also needs to be separated from between-category variation. Of previous studies, only Larochelle and Pineau used a statistical procedure to achieve this separation. Finally, the measurement of categorization time needs to be arranged in such a way as to minimize the effects of either strategic guessing or the retrieval of earlier decisions that render the task less reflective of the underlying structure of semantic memory. If categorization time studies are to tell us anything about the structure of semantic memory and the processes of categorizing concept classes, then great care is needed to avoid guessing strategies or other unintended effects. For example, if only nine categories are used (as in Chumbley, 1986), and these are repeated multiple times, then in the condition where the instance precedes the category, the subject is very likely to develop a strategy of simply generating the appropriate category from memory and then judging whether this is the word that appears on the screen. Use of such a strategy is likely to show measures of the associative strength of the instance–category link to be the best predictor of response time (as Chumbley found).

In this paper, I have two aims. The first is to report an experiment in which many of the potential problems identified above with the regression technique were addressed. This experiment provides a means of clarifying the inconsistencies between Chumbley's results and those of the other researchers. The second aim is to report a second experiment in which the degree to which people rely on associative retrieval versus similarity-based categorization processes was experimentally manipulated. Experimental manipulation of the task is potentially a much more powerful means of identifying the underlying processes than is the purely statistical method of multiple regression analysis.

EXPERIMENT 1

In order to overcome some or most of the difficulties with earlier regression studies, Experiment 1 used the category norms for typicality and PF collected by Hampton and Gardiner (1983), based on the same subject population as used in the present study.[2] These norms provide a large and representative sample of the available category members in each of 12 categories, permitting adequate generalization both within and across categories. The large instance sample sizes allowed regression analyses to be run for each category separately. Categorization time was measured by presenting each category name first, followed by a randomly ordered list of instances and noninstances presented one at a time in a blocked fashion. This procedure reduces the likelihood that subjects are trying to guess the stimuli in advance (the chance of a correct guess would be about 1%) and removes the random variance in decision time due to reading a new category name on each trial. Each instance was presented once only, so that there would be no repetition priming. Under these conditions, it was hoped that categorization time would be a more valid indicator of the relevant interitem differences within each category.

Experiment 1 aimed first to confirm that PF and typicality are separable aspects of semantic memory structure by measuring their independent contributions to predicting categorization time and error rates. By taking PF and typicality as indices of (1) association-based retrieval of prestored "is a" relations and (2) similarity comparison processes respectively, the contribution of these processes to the overall within-category variance in categorization time and response rate can then also be compared, thus addressing the issues raised by Smith (1978) and by the more recent studies. If Casey (1992) and Larochelle and Pineau (1994) are correct in their critique of Chumbley's (1986) results, then there should be substantial effects of typicality in the task, over and above the effects of associative PF.

In performing the regression analyses, a secondary hypothesis was also tested. McCloskey (1980) suggested that effects previously attributed to variations in typicality (or PF) may be owing to a confounding of typicality and PF with familiarity. Clearly, if this was the case, then variance in categorization time would be explainable by a much more general and, hence, less interesting factor, and the task would not reflect anything specifically interesting about semantic memory itself. Hampton and Gardiner (1983) also obtained ratings of item familiarity for their category materials. By including mean rated familiarity for each item in the analysis, McCloskey's suggestion can be rigorously tested in the case of the categorization times measured here. (Other effects of familiarity were reported by Glass & Meany, 1978, Larochelle & Pineau, 1994, and Malt & Smith, 1982.)

Finally, regression analysis was also applied to the correct response rates (or more specifically to the probability of a positive category decision)[3] for individual instances in the 12 categories. Negative responses resulting from a failure to retrieve an "is a" link may be expected to be associated with low PF, whereas those owing to low featural similarity should be associated with low typicality. This analysis therefore provides further information about how the task is performed and, in particular, about the causes underlying a "no" response to a putative category member. Previous research (Chumbley, 1986; Larochelle & Pineau, 1994) has not considered correct response rates as a possible source of converging evidence. Regressions performed on response rates therefore provide a second and important test of the independence of the two dimensions of semantic memory.

Method

Subjects. Sixty volunteers were paid £3 to act as subjects. They were all students at City University London. None had taken part in the Hampton and Gardiner (1983) study.

Design. The 12 categories used by Hampton and Gardiner (1983) were divided into two sets of 6, minimizing the apparent similarity between categories within each set. Each subject categorized lists of words for one of the sets. Each list was presented as a block with items randomized for each subject within blocks, and the order of lists was balanced across subjects. Mean response times (RTs) were calculated across subjects for positive responses to each item in each category.

Materials. All of the words listed in the norms were used. Full details of how the norms were created and the actual words used can be found in Hampton and Gardiner (1983). Briefly, three groups of subjects were employed. One group was given 12 category names, and they had to generate as many examples of each category as they could in a fixed time. PF was based on this group. A second group rated a list of 37–55 category members for each category (sampled independently of the category exemplar production task) for typicality on a 6-point scale. A third group rated the same lists of items organized in the same categories for familiarity on a 6-point scale. Instructions pointed out the difference between the dimensions of typicality, familiarity, and frequency of occurrence in order to help subjects to focus attention on the relevant dimension. Although not part of "standard" typicality or familiarity instructions, this aspect of the Hampton and Gardiner study is advantageous in that it should help to reduce the confounding of the measures and so emphasize their distinctive contributions to categorization. Where appropriate, the subjects had the opportunity of saying that any word was either not a member of the category (in the typicality rating task) or was unknown to them (in either the typicality or the familiarity rating task). Reliability for the three measures was high, averaging .92 within each category, and (crucially for the present purposes) was at the same level for each measure. There were a total of 531 category members used, spread across 12 categories.

To provide negative examples, three additional categories were chosen from Battig and Montague's (1969) norms for each of the 12 categories. Of these three, one was related, one was slightly related, and one was unrelated to the target category. Relatedness of false categories was taken from data published by Herrmann, Shoben, Klun, and Smith (1975), who had subjects perform a clustering-by-similarity task on the 56 categories used by Battig and Montague. For example, for the category Clothing, the related false items were from the category Footwear, whereas for Food Flavorings, the related false items were drawn from the Alcoholic Beverages category. False items were chosen so that, overall, there would be an equal number of expected "yes" and "no" responses for each list and equal numbers of items from each of the three false categories. Since the number of positive items varied between categories, the final lists contained 68–110 words.

Procedure. The subjects sat in front of the display screen of a Commodore CBM 3032 computer, on which the words were displayed. They were told that they would see six lists of words. A category name appeared at the start of each list, in the form of a question, such as "Are the following types of SPORT?" The category name then remained on the screen in the corner of the display, as a reminder. There then followed, one by one, the list of positive and negative items, in a new random order for each subject. The subject pressed one of two response keys as rapidly as possible, to indicate whether or not each item belonged in the named category. After completing each list, the subjects were given a 2-min rest. Instructions emphasized the importance of making as few errors as possible. The whole session took about 45 min.

Results

To remove the undue effect of extreme RTs, 15 latencies (0.1%) of less than 250 msec were excluded from the analysis of mean correct "yes" RTs, and 33 latencies (0.2%) of over 3,000 msec were truncated to 3,000 msec.[4] Mean categorization times for true and false items were obtained by averaging times for correct responses to each item across subjects.

Times taken for correct rejection of false items showed the standard effect of relatedness of negative items (Hampton, 1979; Schaeffer & Wallace, 1970), with mean times of 698 msec for unrelated category items, 795 msec for slightly related items, and 798 msec for strongly related items. Mean true categorization time across all categories was intermediate between these levels at 762 msec, and it varied across categories from 696 msec for Birds to 880 msec for Insects. However, within each category, mean true categorization time for individual instances varied widely, from a low of 600 msec to a high of 2,000 msec. It is this variance that the experiment aimed to predict from the earlier measures of typicality, familiarity, and PF. A split-half reliability measure was obtained for the categorization time data within each category list, by correlating the item means based on the first set and the second set of 15 subjects judging each category list. Corrected reliabilities varied from .63 for Sports to .88 for

Table 1
Correlations Between the Dependent Variable Categorization Time and Log Production Frequency, Typicality, Familiarity, Word Frequency, and Word Length in Experiment 1

	Categorization time with:				
	PF	TYP	FAM	WF	LEN
Birds	−.51	.57	.49	.17	−.07
Clothing	−.66	.56	.81	−.34	.25
Fish	−.69	.79	.61	−.05	−.05
Flowers	−.60	.72	.73	−.22	−.02
Food Flavorings	−.66	.60	.62	−.25	.34
Fruit	−.67	.73	.65	.13	−.09
Furniture	−.74	.84	.35	−.28	.49
Insects	−.74	.72	.66	−.21	.36
Sports	−.54	.54	.32	.17	.12
Vegetables	−.61	.67	.67	−.23	.09
Vehicles	−.64	.62	.44	−.29	.17
Weapons	−.69	.55	.41	−.38	.41
M	−.65	.66	.56	−.15	.17

Note—PF = log production frequency; TYP = typicality; FAM = familiarity; WF = word frequency; LEN = word length.

Fruit, with a mean of .78 (all values were significant, $p < .001$).

Regression analysis.[5] Following Hampton and Gardiner (1983), PF was transformed to log(PF+1) to correct for the skewness of its distribution, which would reduce the linear correlation with categorization time. Scatterplots confirmed that log(PF+1), typicality, and familiarity had essentially linear relations with categorization time. For ease of presentation, all following references to PF refer to the log-transformed variable.

Table 1 shows the Pearson correlations between categorization time and each of the three main independent variables plus two other variables that may be expected to affect RT—word frequency taken from Kučera and Francis (1967), and word length, defined as the number of letters in a word. (These two lexical variables in fact showed little consistent correlation with categorization time.)

The two variables of greatest theoretical interest, typicality and PF, were equally well correlated with categorization time overall, at .66 and −.65, respectively. Familiarity was less well correlated with categorization time (average .56), although for 2 categories, Clothing and Flowers, familiarity had the highest correlation.

Table 2 gives the standardized regression coefficients (β) for the regressions predicting categorization time from seven variables: PF, typicality, familiarity, word frequency and length (as defined previously), *unknown* (which was defined as the number of subjects in Hampton and Gardiner's, 1983, study who judged an item to be unknown to them when rating either typicality or familiarity of items), and *ambiguity* (which was a binary variable defined as 1 if a word had an alternative meaning in the dictionary—e.g., *bass* or *perch*—and zero otherwise). Different methods of achieving the optimal regression solution were tried, with largely similar results and the same conclusions. Table 2 shows the result of removing from the full regression equation in a stepwise fashion any variables entered with the wrong sign[6] or with a nonsignificant regression weight ($\alpha = .05$, one-tailed). (Only three variables entered with the wrong sign—PF for Flowers, and word frequency for Birds and Fruit.)

Comparing the different independent variables, typicality entered 10 of the 12 equations, PF and unknown entered 6 apiece, length was in 3, and ambiguity and familiarity entered only 2 equations. Word frequency did not enter any equations, perhaps because of the constrained nature of the task context (see Becker, 1979). Most importantly, when each category was tested to see whether removing either variable from the full equation led to a significant reduction in R^2, 4 categories identified typicality as a significant predictor, and 2 picked out PF.

The same general pattern of weights emerged when all categories were analyzed together. Dummy variables were entered first to equate for differences in mean categorization time for the different category lists (see Larochelle & Pineau, 1994). Three categories were significantly slower on average than the rest: Insects (144 msec slower), Furniture (84 msec slower), and Food flavorings (82 msec slower). Subsequent forward steps then included the following variables (with associated β weights in the final equation): typicality (.39), unknown (.28), PF (−.21), and length (.10). Multiple R for the final equation was .791, corresponding to 63% of the variance in mean categorization time, of which some 10% could be attributed to the between-category dummy variables. Specific tests for removal of typicality and PF showed that both measures contributed significantly to the variance explained. Typicality contributed an extra 5.6% to the variance explained [$F(1,520) = 77.7$, $p < 10^{-14}$], and PF contributed an extra 1.3% [$F(1,520) = 18.59$, $p < .00002$].

The results of the full analysis show that four factors contributed to categorization time: typicality, PF, word length, and the probability of an item being unknown. More concretely and as a means of comparing the relative effect sizes, going from highest to lowest possible values on each scale increased categorization time by 292 msec for typicality and 103 msec for PF, while each extra percent of the subjects not knowing an item increased categorization time by 6 msec, and each letter of a word took an extra 7 msec to process. Interestingly, there was little evidence that rated familiarity affected categorization time in the present task, once the effect of unknown items was removed. McCloskey's (1980) concerns about the familiarity confound in typicality ratings may then be restricted to cases where items are so unfamiliar as to be unknown to some subjects.

Finally, the level of prediction achieved, multiple R, corresponded closely to the reliability measures for categorization time across different categories, both in mean levels ($R = .765$, reliability = .778) and in the correlation across categories ($r = .77$, $n = 12$, $p < .005$). The close match suggests that reliability level for categorization time was probably a limiting factor restricting the level of fit achieved in the regression equations.

Response probability. A second set of regression analyses were used to predict the proportion of "yes" responses for each category member from the five variables:

Table 2
Standardized Regression Weights (β) for Each of the Significant Predictors of Categorization Time in Experiment 1, and *R* for the Optimal Regression Equation

Category	TYP	PF	FAM	WF	LEN	UNK	AMB	*R*
Birds	.44	–	–	–	–	.34	–	.648
Clothing	–	−.32	.46	–	–	.24	–	.852
Fish	.79	–	–	–	–	–	–	.788
Flowers	.41	–	–	–	–	.43	–	.775
Food flavorings	–	−.36	–	–	.18	.54	–	.830
Fruit	.53	–	.37	–	–	–	–	.791
Furniture	.84	–	–	–	–	–	–	.840
Insects	.31	−.30	–	–	.22	.28	–	.837
Sports	.30	−.33	–	–	–	–	.21	.614
Vegetables	.45	–	–	–	.22	.50	.18	.819
Vehicles	.33	−.41	–	–	–	–	–	.681
Weapons	.24	−.55	–	–	–	–	–	.715
All categories:	.386	−.206	–	–	.099	.279	–	.791

Note—TYP = typicality; PF = log production frequency; FAM = familiarity; WF = word frequency; LEN = word length; UNK = unknown; AMB = ambiguity.

PF (log-transformed), typicality, familiarity, unknown, and ambiguity, all as defined previously. To disconfound the typicality scale from the proportion of subjects rejecting a category exemplar, mean typicality values were recalculated from the norms for this analysis by excluding any subjects who gave a rating of 6 (= *not in the category*). The mean typicality values thus reflected the mean typicality judgment of those subjects who believed that the item was a category member.

Across all categories, typicality was much the best predictor of response rate (β = .538), with unknown (.332) and ambiguity (.129) also predictive. Multiple *R* was .692. PF and familiarity had no predictive value. When regressions were calculated for each individual category separately, in no case did PF enter significantly. We can therefore conclude that, apart from unknown and ambiguous items, there was only one reason for the subjects to reject an ostensible category member—its low typicality. In no case did failure to retrieve a category link, as indexed by low PF, appear to have led to negative responses.

Discussion

As expected from earlier studies, item differences in mean categorization time proved to be highly predictable from measures of category instance gradedness. For the two variables of theoretical interest, typicality and PF, the results supported the hypothesis that the two variables reflect partly independent sources of variance in categorization time. Each variable made an independent contribution to the prediction of categorization time. By contrast, the second dependent variable, the probability of a "yes" response, was predicted entirely by typicality, without any independent contribution from PF. The results therefore support the regression studies of Casey (1992) and Larochelle and Pineau (1994), who found typicality to be a consistent predictor of categorization time, and suggest that Chumbley's (1986) results were unrepresentative. The results also go beyond previous research in a number of ways. First, the validity of the measurement of categorization time was improved by employing a procedure using single presentation of many instances per category and a listwise presentation in order to reduce strategic guessing, repetition priming, and possible sampling bias effects. Second, the use of response rate as a secondary dependent variable provided converging evidence of the separate effects of PF and typicality in the task.

The independent effects of the two variables suggest that no single process model, involving simply the retrieval of prestored "is a" relations in an associative network, nor just the comparison of feature overlap can account fully for the time taken to categorize words. This conclusion can be made on the basis of the present data, without concern for the generality of the results to other versions of the task, and supports a similar conclusion reached by Larochelle and Pineau (1994). The association of positive response rate with typicality alone provides strong evidence that categorization involves more than the retrieval of a prestored category relation (as proposed, for example, by Chumbley, 1986). It suggests that "no" responses arise to putative category members only when atypical instances fail to reach a sufficient degree of similarity to match the criterion for inclusion in the category (Hampton, 1979; McCloskey & Glucksberg, 1979). In effect, even in a speeded decision task, category membership appears to be dictated solely by semantic content and not by association strength. The fact that PF affected categorization time without affecting categorization response probability suggests that rapid retrieval of an instance–category "is a" relation may have been used as a means of deciding that an item belonged in the category, but that failure to retrieve such a relation was not used as a means of deciding that the item did not belong. Retrieval of an "is a" relation is a sufficient, but not necessary, basis for making a "yes" response.

The results of Experiment 1 support a model of semantic memory categorization in which both retrieval of "is a" links and feature comparison processes contribute (in varying degrees) to the overall variance in catego-

rization time. Lorch (1978, 1981) also argued for a mixed model on the basis of finding independent effects of accessibility and similarity on false categorization sentences. Collins and Loftus's (1975) spreading activation model involved not only retrieval of prestored category links from a semantic network but also a variety of additional routines for computing a categorization decision in other less direct ways. However, there is little or no direct experimental evidence for the two processes acting on true categorization responses. The approach adopted in Experiment 2 was therefore to seek experimental manipulations of the categorization task that may be expected to have differential effects on the influence of the two variables on categorization RT and response rate. Experimental dissociation of the effects of the variables would constitute much stronger evidence for the mixed model of categorization. Experiment 2 also used category materials selected in such a way as to manipulate the two variables in a controlled quasi-experimental design. Items were selected to provide separate measures of the effects of PF and typicality, and experimental manipulations were chosen that, it was predicted, would dissociate the two variables by showing different effects on the relation between each variable and the dependent measures of categorization time and response rate.

EXPERIMENT 2

The aim of Experiment 2 was to discover whether the effects of typicality and PF on categorization time and response rate would interact differentially with manipulations of the experimental task context. Experiment 1 showed that the two variables had independent effects on categorization time and that response probability was associated with typicality and not at all with PF. The logic of Experiment 2 was to find two different manipulations of the task. One manipulation was designed to modulate the effects of typicality on categorization, while leaving the effects of PF unchanged. The second manipulation was designed to achieve the reverse dissociation, interacting with the PF effect but leaving the typicality effect unchanged.

For the first of these manipulations, the difficulty of discriminating positive from negative category instances was varied. Varying the task difficulty in this way should lead a subject to set a higher decision criterion in the feature comparison process. For example, according to McCloskey and Glucksberg's (1979) property comparison model, more evidence of the degree of feature overlap would need to be sampled before responding, in order to maintain a reasonable level of accuracy. According to the present assumptions, this slowing up of the feature comparison process should affect the size of the typicality effect, while leaving the PF effect unchanged. A high-PF instance is still likely to be categorized through the retrieval of a strong instance–category "is a" link. However, in the absence of a strong category association, the feature comparison decision process should be differentially slowed more for atypical instances than for typical instances.

The specific manipulation used in the experiment was taken from the study by McCloskey and Glucksberg (1979). They showed that if the false item–category pairs in a list to be categorized were all unrelated, then true RTs were both faster and less sensitive to differences between typical and atypical category members. When the relatedness of false item–category pairs was increased, then the criterion for the accumulation of sufficient evidence to make a positive decision became more strict, with a resulting increase in the difference in categorization time between typical and atypical category members. McCloskey and Glucksberg, in fact, used PF as the basis for selecting high- and low-typicality instances for their categories so that their instances differed in both PF and typicality. For the present experiment, the strong prediction can be made that their result should be found for materials that differ in typicality but should not be found for materials that differ only in PF. Experiment 2 also answered a potential criticism of McCloskey and Glucksberg's results. They showed an increased typicality effect for the condition that included related false items, but this increase was found in the context of a general slowing down of all RTs and could therefore have merely reflected the skewed distribution of RTs in general. Since Experiment 2 predicts that the increase will occur specifically for differences in typicality and not for differences in PF, this general interpretation of their result would be ruled out by the predicted pattern of results.

An earlier study by the author (Hampton, 1988) showed, as predicted, that introducing related false items into a list of true instance–category pairs increased the typicality effect on categorization times from 18 to 48 msec, but it did not increase the PF effect (40 vs. 37 msec). Two related false conditions were used: one in which all nonmembers were related, and a second in which only half the nonmembers were related. Both led to an increase in the typicality effect; however, in the all-related condition (corresponding to McCloskey & Glucksberg's, 1979, Experiment 2), there were some subjects who apparently adopted a different strategy for doing the task. These subjects showed an increase in the PF effect on categorization time and a much higher false positive error rate, suggesting that they could have been responding "yes" on the basis of finding any semantic association between the item and the category, regardless of whether the item really was a category member. For Experiment 2, therefore, the false items included some related items and some unrelated items. In an attempt to increase the effectiveness of the manipulation, false-item relatedness in Experiment 2 was deliberately confounded with instructions to subjects, which either encouraged speed (in the unrelated false condition) or advised caution (in the condition with related false items). Instructions to concentrate on accuracy of responding should also discourage the undifferentiated association strategy just described.

The second manipulation introduced in Experiment 2 was designed to produce a reverse dissociation by differentially affecting the PF effect on categorization. There

was no obvious manipulation in the literature corresponding to the McCloskey and Glucksberg manipulation of false-item relatedness interacting with typicality, which could be expected a priori to influence the retrieval of instance–category links. A manipulation was therefore chosen by analogy with an effect in the lexical decision task literature. Scarborough, Cortese, and Scarborough (1977) found that the normal word frequency effect on lexical decision time (that high-frequency words are more rapidly verified as words than are low-frequency words) was attenuated if the words were primed by having been read earlier in the experiment. Repetition priming, therefore, appears to reduce or even remove the standard frequency effect. By analogy, a priming manipulation was introduced into Experiment 2, with the intention that it should reduce the difference in categorization time between high- and low-PF instances. Retrieving the meaning of the instances in an earlier semantic decision was expected to leave their associative category links in an activated state and, hence, to attenuate the difference between high- and low-PF instances.

The priming task required subjects to categorize items with respect to a more superordinate category. For example, if an instance–category pair were *swift*–Bird, then, in the priming phase of the experiment, a subject would be asked to judge the instance–category pair *swift*–Creature. Later, in the main part of the experiment, the subject would then judge the pair *swift*–Bird. The expectation was that this form of repetition priming should work to prime category relations for the repeated words. The low-PF instances should therefore show greater priming than the high-PF instances, since the latter would already have easily accessible instance–category links. Since the prior exposure did not directly involve categorization of the item in the target category, it was predicted that the typicality effect would remain unaffected by this priming manipulation. Deciding, for example, that an atypical instance like *penguin* is a Creature does not necessarily make it any easier to decide later that a *penguin* is a Bird. However, deciding that a low-PF instance, such as *cuckoo* is a Creature may be expected to facilitate a later decision that it is a Bird.

Since this prediction is the converse of that derived for the manipulation of false-item relatedness, by including both manipulations in a single design, it was hoped to show a double dissociation of typicality and PF effects within the same experiment. In order to provide independent measures of the typicality and PF effects, it was necessary to select appropriately controlled sets of materials. It proved difficult to select a fully orthogonal set according to a 2×2 design of high and low typicality with high and low PF, largely because the low–low set of words tended to be more unfamiliar than the rest. As an alternative, two sets of materials were designed to be used on different subject groups. The first set maximized the manipulation of typicality between two sets of instance–category pairs, while holding PF constant. The second set maximized the difference between sets in their PF, while holding typicality constant. Mean familiarity was also held constant across all item sets.

Method

Subjects. The subjects were 96 undergraduate student volunteers at The City University, London, who were paid £3 to take part. They were assigned on order of appearance at the laboratory into four equal groups of 24 subjects each.

Design. The design incorporated two between-group factors. The first was measure (typicality vs. PF). In order to increase the difference between high and low values on each of the typicality and PF measures, the two measures were manipulated for different groups of subjects. That is, half the subjects took part in conditions considering effects of typicality on categorization time, and half took part in conditions considering effects of PF on categorization time. These two halves of the experiment were identical in every respect, except for the materials used for the true instance–category pairs. Dividing up the materials effects between subjects in this way enabled a larger difference between high and low items to be achieved on each measure, subject to the same balancing considerations as before. The second between-group factor was criterion. Two manipulations were deliberately confounded in order to produce a strong manipulation of the subjects' decision criterion for making a categorization response. In one set of conditions, the subjects were told that false items would be easy to reject, and they were encouraged to proceed as fast as they could, without making too many errors. Speed was again emphasized at the end of the instructions, and the false items in the list were in fact all unrelated to their paired categories. The other half of the subjects were told (truthfully) that some of the false items would be difficult to decide about, and they were warned to go carefully while still responding as fast as was consistent with few errors. Accuracy was again mentioned at the end of the instructions; in the subsequent task, 60% of the false instance–category pairs were indeed related. To summarize, there were four groups of subjects taking part in four conditions, which will be referred to as follows: *typicality–speed*, *typicality–accuracy*, *PF–speed*, and *PF–accuracy*, where *speed* refers to a low-criterion condition with speed instructions and unrelated false items, and *accuracy* refers to a high-criterion condition with accuracy instructions and 60% related false items.

In addition to these between-group factors, there were also two within-subject factors. The first was the centrality of a true item (where *centrality* is used as a general term to refer either to typicality or to PF). Half the true items to be judged were high (on typicality or PF, depending on the condition), and half were low. The second factor was priming. Half the words seen in the critical test session had been seen earlier in a priming session, paired with a more superordinate category name. The remaining half were unprimed and were seen for the first time in the experiment at test.

Materials were fully balanced across priming condition, so that the full design involved multiples of eight subjects.

Materials. Two sets of materials were devised with some overlap between them. All measures were based on the Hampton and Gardiner (1983) category norms. One set (the typicality set, used for typicality conditions) was composed of 32 high-typicality (mean typicality = 1.42) and 32 low-typicality (mean typicality = 2.93) instance–category pairs, which were chosen to have matched PF (mean PF = 13.8 and 13.5, respectively) and matched familiarity (mean familiarity = 1.55 and 1.52, respectively). The other set (the PF set) contained 32 high-PF and 32 low-PF instance–category pairs (mean PF = 33.6 and 4.5, respectively), matched for typicality (1.85 and 1.86, respectively) and familiarity (1.45 and 1.53, respectively). The pairs were taken from all 12 categories in Hampton and Gardiner and are listed in Appendix A. Each category always occurred equally often with high items and with low items across different conditions. The initial priming session consisted of a cat-

egorization task, similar to that used for the main test. Sixteen of the 32 high items and 16 of the 32 low items for the particular set of pairs for the condition were used in the priming session, paired with one of the following categories to give a true instance–category pair: Creatures, Man-Made Objects, Plants, Recreations, or Food. In addition to the 32 true pairs in the priming session, there were 32 false pairs. These false pairs were composed of words that would appear as false items later in the main test session. In the speed condition, all of these false pairs were unrelated items paired with one of the same five general superordinates (e.g., *copper*–Recreation). In the accuracy condition, 20 of these 32 false pairs were related in meaning to the category name to be used later (e.g., *bat*–Bird) but were not necessarily related in meaning to the more superordinate term used in the priming session (*bat*–Food). Finally, there were 10 practice items at the start of the list, and there were eight filler items that, while true for the priming session, would be false pairs for the test session, in order to discourage the subjects from using the response made in the priming session as a way of predicting the response in the test. With 10 practice trials, 32 true trials (16 high and 16 low), 32 false trials, and 8 fillers trials, the priming session comprised 82 trials in all.

After the priming session and a short break, there followed the test session. The list of items for the test session contained all the true and false items from the priming session, but now paired with the original 12 categories. In addition, the remaining 32 true pairs were included, as unprimed high (16 items) and unprimed low (16 items) pairs. Likewise, there were 32 new unprimed false items. In the speed condition, all false pairs were unrelated. In the accuracy condition, 60% of both primed and unprimed pairs were semantically related. To construct false pairs, the same 12 categories were used with roughly the same relative frequency as used for true pairs. Related false items were from neighboring categories or were potentially borderline cases of the category. Unrelated false items were chosen from categories such as Cities, Countries, Toys, and Musical Instruments. Illustrative examples are shown in Appendix B. The final list was completed with 12 new practice items to introduce each of the new category terms (practice items included true items and related or unrelated false pairs depending on the condition) and with the eight fillers from the priming session that, having been true before, were now falsely paired with categories. There were 148 trials in all.

Procedure and Apparatus. The apparatus was the same as in Experiment 1. The subjects were given one of two written sheets of instructions, according to whether they were in the speed condition or the accuracy condition. They were then shown how to start the sequence of trials for the priming session by pressing one of the response keys. The first 10 trials were discounted as warm-up trials, but the subjects were not aware of this and simply carried straight on. The order of critical instance–category pairs was randomized for each subject. Each pair was individually presented in the center of the display screen in uppercase, with the instance displayed simultaneously with, and directly above, the category name. A warning asterisk signaled the start of each trial. The pair remained on the screen until the subject had responded by pressing one of two keys, one for each hand. The "yes" key was placed by the subject's preferred hand. After 41 trials, the subjects were given a break and continued in their own time when they were ready. After the end of the priming session, there was a short break while preliminary results were printed out, and a second program was loaded into the computer. The main test session then followed. Again, the first 12 trials were discounted as warm-up trials. Each category name occurred once during these 12 trials. There was a break halfway through the session. After the experiment, the subjects were debriefed and asked to tell of any ambiguous items or other problems they may have encountered.

Results

Latencies less than 250 msec were excluded, and latencies over 3,000 msec (less than 1%) were truncated to 3,000 msec. (An alternative analysis was run in which long latencies were excluded if greater than 3 standard deviations above the mean calculated separately for each subject, with the same general results.) Table 3 shows the full set of mean categorization times for each condition in the experiment.

A five-way ANOVA was conducted of the complete design, with factors of centrality (high vs. low values of either PF or typicality), measure (PF vs. typicality), priming (primed vs. unprimed), criterion (speed vs. accuracy instructions confounded with false-item relatedness), and word set (words were balanced between primed and unprimed conditions). The following effects were significant on a min F' test: main effects of centrality [min $F'(1,139) = 6.21$, $p < .01$], priming [min $F'(1,205) = 32.6$, $p < .001$], criterion [min $F'(1,131) = 34.7$, $p < .001$], and a two-way interaction between centrality and criterion [min $F'(1,198) = 4.14$, $p < .05$). Items that were high typicality or PF were faster than those that were low. Primed items were faster than unprimed. The subjects reacted faster in the speed instruction condition where the criterion was lower and false items were unrelated. The interaction reflected the fact that centrality effects were greater in the accuracy conditions than in the speed conditions.

The significance of the main effects of priming and of criterion indicate that the experimental manipulations were indeed affecting the categorization task.

Because of the complexity of the design, further analysis of the results focused on testing particular hypotheses concerning the effects of criterion and priming on the typicality and PF effects.

Effects of criterion. The manipulation of criterion was clearly very effective, resulting in categorization times that differed overall by some 250 msec. The prediction made for this manipulation was that a more cautious criterion should increase the typicality effect, while not affecting the PF effect. Table 3 shows the effects of the factor under the different experimental conditions. For the typicality comparison condition, the effect was exactly as predicted. Under speed conditions, the typicality effect was 22 msec (primed) and 23 msec (unprimed). Under accuracy conditions, it rose to 100 msec (primed) and 114 msec (unprimed). A four-way analysis of variance (ANOVA) of the typicality comparison condition, with criterion, priming, word set, and typicality as factors, showed clearly significant main effects of criterion [min $F'(1,64) = 13.16$, $p < .001$] and priming [min $F'(1,100) = 14.17$, $p < .001$] and a marginal effect of typicality [min $F'(1,67) = 3.57$, $p < .10$]. Most importantly, there was also a significant interaction between criterion and typicality [min $F'(1,104) = 3.91$, $p = .05$]. There was no interaction at all between typicality and priming [$F < 1$, by both subjects and items analyses]. It is clear that the typicality effect responds strongly to changes in criterion, but not at all to priming. The predictions were therefore fully supported.

For the PF comparison condition, a different pattern was seen. For unprimed category–instance pairs, the PF effect was 42 msec for speed conditions and 57 msec for accuracy conditions. Thus, the typicality effect for un-

Table 3
Mean Categorization Times (in Milliseconds), *SD*s, and Percentage Error (PE) Rates for True Items in Each Condition in Experiment 2

	Priming Condition						
	Primed			Unprimed			Priming
	M	*SD*	PE	*M*	*SD*	PE	Effect
Speed Condition							
Typicality							
High	754	197	4	808	217	6	54
Low	776	212	4	831	223	6	55
Typicality Effect	22			23			
PF							
High	695	104	3	730	106	4	35
Low	696	95	5	772	108	7	76
PF Effect	1			42			
Accuracy Condition							
Typicality							
High	961	231	3	1,014	231	3	53
Low	1,061	290	7	1,128	327	8	67
Typicality Effect	100			114			
PF							
High	925	224	4	1,027	307	4	102
Low	979	269	4	1,084	304	4	105
PF Effect	54			57			

primed pairs increased by some 91 msec as criterion was manipulated, whereas the PF effect increased by only 15 msec. Given that, in the earlier study by Hampton (1988), the PF effect actually decreased slightly as the task became harder, it is likely that the PF effect is unaffected in any significant way by changes in criterion when there is no priming. For the primed condition, the manipulation of criterion did increase the PF effect. This increase was observed because priming removed the PF effect, but only in the speed condition. The effects of priming are discussed in more detail in the following section.

Effects of priming. Priming was the second manipulation introduced in this experiment, and it was predicted to interact with PF but not with typicality. Priming effects were defined as the difference in mean categorization time between primed and unprimed pairs. Tables 3 and 4 show that all types of pair (both true and false) showed positive priming in all four subject groups. The manipulation was therefore clearly effective, speeding categorization time by some 35–105 msec. For three of the between-subject conditions, the effects of priming did not, however, interact with either the typicality or the PF of instance–category pairs. In both typicality comparison conditions, the priming effects were equivalent for high-typicality pairs (54 and 53 msec) and for low-typicality pairs (55 and 67 msec). In the PF–accuracy condition, priming was greater but again did not differentiate between high-PF (102 msec) and low-PF (105 msec) pairs. For the PF–speed condition, however, there was an interaction between priming and PF. Priming was stronger for low-PF (76 msec) than for high-PF (35 msec) pairs. Put another way, the PF effect (42 msec) observed for unprimed pairs was completely removed (1 msec) when pairs were primed in the speed condition. A three-way ANOVA of the PF–speed condition, with priming, word set, and PF as factors, confirmed a significant interaction between priming and PF with subjects as random factor [$F(1,22) = 5.17$, $MS_e = 1{,}893$, $p < .05$] and with words as random factor [$F(1,60) = 4.03$, $MS_e = 2{,}345$, $p < .05$].

Within the speed condition, where false items were all unrelated, there was, therefore, support for the prediction that priming would differentially speed access to the low-PF items. Priming was strongest for low-PF instances (76 msec), intermediate for the high- and low-typicality instances that had medium to low PF (54 and 55 msec), and least for high-PF instances (35 msec). The priming factor, therefore, dissociated PF and typicality as measures, in that typicality showed zero interaction with priming under both speed and accuracy conditions, whereas PF showed the predicted interaction in the speed condition.

Contrary to expectation, the results did not show an interaction between priming and PF in the PF–accuracy condition. In this condition, the priming effects were larger and of equivalent size (100 msec) for both high- and low-PF instances. This interaction between the two major manipulations of the experiment was unexpected. Given that the subjects were responding more cautiously in general in the accuracy condition, the increased size of the priming effect might have been the result of the need to access a greater amount of relevant semantic information of all kinds (including property and category associations), with a corresponding increase in the importance of recent access to the word's meaning. It appears that priming only helped low-PF instances differentially in the condition where any semantic similarity is sufficient for a categorization response—namely, the speed condition. Where greater discrimination was needed in the

Table 4
Mean RTs (in Milliseconds) and *SD*s for False Items in Experiment 2

	Priming Condition				
	Primed		Unprimed		Priming
	M	*SD*	*M*	*SD*	Effect
Speed Condition					
Unrelated False					
Typicality	788	191	830	211	42
PF	743	110	784	136	41
M	765		807		42
Accuracy Condition					
Related False					
Typicality	1,336	319	1,402	369	66
PF	1,292	391	1,328	382	36
M	1,314		1,365		51
Unrelated False					
Typicality	1,025	256	1,054	269	29
PF	1,020	282	1,099	343	79
M	1,023		1,076		53

accuracy condition, priming helped both high- and low-PF instances equally.

False items. RTs for correctly rejecting false items are shown in Table 4. Since the same false-item materials were used in both typicality and PF comparison conditions, they have been averaged in the table. First, as expected, unrelated false items (1,023 and 1,076 msec) were faster than related false items (1,314 and 1,365 msec) in the accuracy condition. These same unrelated false items were again much faster (765 and 807 msec) in the speed condition, where no related false items were present. In addition, Table 4 shows a consistent priming effect in all three sets of means of 42–53 msec. Since this priming effect was unaffected by the relatedness of the false items (compare effects of 51 msec for related false and 53 msec for unrelated false in the accuracy condition), it is likely that priming reflects the speeding of some process that is occurring prior to the decision stage. This conclusion is strengthened by the similar priming effects shown by true items. Neither typicality of true items nor relatedness of false items showed any interaction with the priming manipulation.

Errors. Error rates for true responses are shown in Table 3. They were subjected to a five-way ANOVA, with centrality (high vs. low), measure (typicality vs. PF), priming, criterion (speed vs. accuracy), and word set as factors. The results were very clear. Only two effects were significant across both words and subjects: centrality [across subjects, $F(1,88) = 8.88$, $p < .005$; across words, $F(1,120) = 4.52$, $p < .05$] and the three-way interaction of centrality, measure, and instructions [min $F'(1,195) = 4.66$, $p < .05$].

The reason for the significant interaction was that the main effect of centrality, with high word pairs giving fewer errors than low, is restricted to two of the four conditions. Low-typical words generated more errors only under the accuracy condition (where related false items made the categorization decision more difficult). Conversely, low-PF words yielded more errors only under the speed condition. The interaction thus provides clear additional support for the functional dissociation of similarity-based and association-based effects in categorization. In the accuracy condition, the results of Experiment 1 were confirmed in that errors were most common for low-typicality instances (mean typicality = 2.93), intermediate for the PF materials (mean typicality = 1.85), and least for high-typicality instances (mean typicality = 1.42). By contrast, in the speed condition, error rates were highest for the low-PF instances (mean PF = 4.5), slightly lower for the typicality materials (mean PF = 13.6), and least for high-PF instances (mean PF = 33.6).

Discussion

The results of Experiment 2 can be summarized as follows. First, changing the relatedness of false items and encouraging the subjects to raise their decision criterion had the effect of slowing down atypical category members more than typical members. High- and low-PF members were equally slowed down by the manipulation in the unprimed condition. The results therefore confirm that the increase in the centrality effect on RT, discovered by McCloskey and Glucksberg (1979), works mainly by slowing down atypical instances, as opposed to low-PF instances. This result is entirely as would be predicted by feature comparison models. Comparing the pattern of errors between the two criterion levels confirms this interpretation. For the high-criterion condition, the pattern of errors followed the results of Experiment 1, with most errors made to atypical instances, fewest made to typical instances, and no effect of PF on error rate. When the criterion was low, however, and it was easy to discriminate between true and false pairs, then no more errors were made to atypical than to typical category members. Instead, errors were more likely to low-PF instances. With the emphasis placed on speed, and no related distractors, the subjects appeared to rely more heavily on a strategy where the retrieval of any semantic association may have in and of itself been sufficient to make a categorization decision. Thus, on occasion, the failure to retrieve any semantic association between item and category may have been used as the basis to erroneously reject a low-PF instance from the category.

The new factor introduced in Experiment 2 was the repetition priming manipulation. Predictions for this manipulation could be based only on an analogy with its interaction with word-frequency effects in lexical decision times and so could not be made with the same degree of confidence. The primary goal, however, was to find a manipulation that would show an interaction with PF effects but would not interact with typicality effects. As such, the manipulation was at least partially successful. For typicality effects, as predicted, priming with an earlier decision that the item was in a more superordinate category was completely ineffective in changing the difference in RT between high- and low-typical instances (i.e., both sets of items were equally primed by the repetition). For PF effects, the priming was effective in re-

moving the difference between the categorization time for high- and low-PF instances, but this occurred only in the low-criterion speed condition. This result is consistent with the suggested strategy for this condition that subjects are using the ease of retrieval of semantic information relevant to the decision as a way of reaching a quick category decision. Activation of the low-PF words in the semantic priming task could be expected to activate relevant connections to the later category term and lead to a more rapid categorization of these items.

For the high-criterion accuracy condition, the priming effect did not interact with PF. Both high- and low-PF instances were primed to the same extent. PF still produced differences in categorization speed, roughly equivalent to those in the unprimed–speed condition. Any explanation of this unexpected result can only be post hoc. The pattern of error data suggests that, in the speed condition, categorization may have been based on undifferentiated semantic associations. This strategy gave rise to more errors for low-PF instances and a priming effect that was greater for low-PF instances than for high-PF instances. In the accuracy condition, however, this strategy would have led to unacceptably high error rates, since many of the related false items would have attracted positive responses. If the priming manipulation simply increased the availability of undifferentiated associations, then its effect in the accuracy condition may have been to increase access to the featural information needed for a category decision based on similarity. Only the high-PF instances, which are more likely to have specific "is a" links to the category, would then be categorized on the basis of specific category associations, whereas the rest would be categorized through a feature-comparison decision process.

GENERAL DISCUSSION

This research has established that internal category structure is graded in more than one sense and that the process of making speeded categorization decisions is sensitive to at least two forms of gradedness: the specific association of an item with a category as a category member, and the similarity or representativeness of an item in the category. Experiment 1 showed that these two kinds of gradedness, as indexed respectively by PF and typicality, can be differentiated through their contribution to within-category variance in the speed and accuracy with which items are categorized. Each variable made a significant independent contribution to a regression predicting RT, whereas only typicality provided a prediction about the probability of a "yes" response. Experiment 2 established a more radical dissociation between the two forms of gradedness by showing quite different patterns of interaction with manipulations of criterion and repetition priming.

Experiment 1 built on earlier research using regression methods and introduced a number of improvements, using an adequate sample of instances per category, using the same population of students for each measure, and using instructions designed to separate out familiarity, frequency, and typicality dimensions. The measurement procedure also avoided repetition of items and guessing strategies that may have had strong effects in earlier experiments (e.g., Chumbley, 1986). The results showed that both PF and typicality made significant contributions to explaining why some category members are categorized more rapidly than others. The experiment also showed, for the first time, a clear dissociation between the two measures in that, of the two, only typicality predicted the likelihood of a "no" response to category members.

Experiment 2 provided further evidence for the dissociation of the two dimensions of semantic memory. The interaction or the relatedness of false items with instance centrality effects, demonstrated by McCloskey and Glucksberg (1979), was shown to be specific to the typicality dimension. The difference in RT between typical and atypical instances was magnified by the increased difficulty of the task, as was the difference in error rates between the two kinds of instances. By contrast, low-PF instances did not show either of these effects relative to high-PF instances in the unprimed condition and, in fact, were less prone to error when related false items were included. The experiment thus clarified McCloskey and Glucksberg's result, showing not only that the effect was not simply an effect of slower RTs across the board but also that the effect works specifically on only one of the centrality dimensions—namely, typicality.

The repetition priming manipulation was introduced by analogy with the word-frequency effect in lexical decision and was predicted to reduce or remove the effect of differences in PF. This prediction was borne out, but only in the speed condition. Priming with a superordinate categorization removed the difference in categorization time for high-PF and low-PF instances in this condition, while leaving differences in categorization time due to typicality unaffected. Error rates supported the dissociation, with more errors made to low-typicality instances in the accuracy condition, but more errors made to low-PF instances in the speed condition. The occurrence of errors to low-PF instances can be taken as clear evidence of the use of an association-based strategy in this condition.

No attempt has been made to use the present data to motivate a process model of the categorization process. It is probable that the cognitive system is too flexible in its processing to warrant such an approach. The experiments presented here show clearly how quite different processes may be involved under different task conditions. It has been argued first that high-PF instances may be categorized on the basis of the retrieval of a strong "is a" category link between the item and the category. PF had a significant effect on categorization times in Experiment 1 and in three of the four conditions of Experiment 2. Second, it appears that there is usually something akin to a feature comparison or similarity computation process involved in categorization decisions. Except in the unusual circumstances where all false items are quite unrelated, the best way to discriminate true from false

items appears to be to retrieve the semantic content of the words' meanings and to use that information as the basis for a categorization. This process is the best way of explaining the robust typicality effects on both RT and error rates seen whenever there are false-related items in the experiment.

The direction taken by this research has been toward a broader exploration of the best ways of devising "clean" measures of memory structure and decision processes and experimental manipulations that produce consistent and comprehensible effects. The present results should be seen as work toward this goal, indicating the need for independent consideration of association- and similarity-based effects within the framework of modeling semantic memory. They also provide a demonstration that manipulation of criterion involves a major shift in the way in which the category verification is performed. Future investigations of this task can use this manipulation to study the associative–retrieval and the similarity–comparison aspects of categorization in relative isolation.

The research has been motivated at a more general level by a distinction between an associationist memory system, in which the operation and structure of the database is determined by frequency of use, and a content-addressable memory system, in which the operating characteristics of the database are determined by the nature of the objects being represented. Integration of these two basic architectures into a common representational system remains an important challenge for cognitive science. The results of the present study of categorization time suggest that semantic memory shows important aspects of both kinds of system.

An interesting corollary of the general distinction of association- and similarity-based structures is to consider the effects of typicality and association strength as "micro" examples of the "macro" cognitive heuristic strategies identified by Tversky and Kahneman (1974) in the judgment of subjective probability. Typicality can clearly be linked to their notion of a representativeness heuristic. In categorization tasks, subjects decide on category membership on the basis of how representative of the category an item appears to be. Production frequency is, on the other hand, easily identified with Tversky and Kahneman's availability heuristic. When responding fast, in an easily discriminated list context, subjects may decide category membership simply on the basis of how easily any semantic link can be found between the item and the category. Manipulation of criterion in the task may therefore lead subjects to set up either availability-based or representativeness-based task-specific strategies.

REFERENCES

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

Anderson, J. R. (1991). A rational analysis of categorization. *Psychological Review*, **98**, 409-429.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **11**, 629-654.

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut Category Norms. *Journal of Experimental Psychology Monograph*, **80** (3, Pt.2).

Becker, C. A. (1979). Semantic context and word frequency effects in visual word recognition. *Journal of Experimental Psychology: Human Perception & Performance*, **5**, 252-259.

Brooks, L. R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 170-211). Hillsdale, NJ: Erlbaum.

Brooks, L. R. (1987). Nonanalytic cognition. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 141-174). Cambridge: Cambridge University Press.

Casey, P. J. (1992). A reexamination of the roles of typicality and category dominance in verifying category membership. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 823-834.

Chang, T. M. (1986). Semantic memory: Facts and models. *Psychological Bulletin*, **99**, 199-220.

Chumbley, J. I. (1986). The roles of typicality, instance dominance, and category dominance in verifying category membership. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **12**, 257-267.

Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review*, **82**, 407-428.

Conrad, C. (1972). Cognitive economy in semantic memory. *Journal of Experimental Psychology*, **92**, 149-154.

Fodor, J. A., & Pylyshyn Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, **28**, 136-196.

Glass, A. L., & Holyoak, K. J. (1975). Alternative conceptions of semantic memory. *Cognition*, **3**, 313-339.

Glass, A. L., & Meany, P. J. (1978). Evidence for two kinds of low-typical instances in a categorization task. *Memory & Cognition*, **6**, 622-628.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning & Verbal Behavior*, **18**, 441-461.

Hampton, J. A. (1984). The verification of category and property statements. *Memory & Cognition*, **12**, 345-354.

Hampton, J. A. (1988, November). *Evidence for a two-process model of categorization*. Paper presented to the annual meeting of the Psychonomic Society, Chicago.

Hampton, J. A. (1993). Prototype models of concept representation. In I. van Mechelen, J. A. Hampton, R. S. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 67-95). London: Academic Press.

Hampton, J. A., & Gardiner, M. M. (1983). Measures of internal category structure: A correlational analysis of normative data. *British Journal of Psychology*, **74**, 491-516.

Herrmann, D. J., Shoben, E. J., Klun, J. R., & Smith, E. E. (1975). Cross-category structure in semantic memory. *Memory & Cognition*, **3**, 591-594.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Larochelle, S., & Pineau, H. (1994). Determinants of response times in the semantic verification task. *Journal of Memory & Language*, **33**, 796-823.

Loftus, E. F. (1973). Category dominance, instance dominance, and categorization time. *Journal of Experimental Psychology*, **97**, 70-74.

Loftus, E. F., & Scheff, R. W. (1971). Categorization norms for fifty representative instances. *Journal of Experimental Psychology*, **91**, 355-364.

Lorch, R. F., Jr. (1978). The role of two types of semantic information in the processing of false sentences. *Journal of Verbal Learning & Verbal Behavior*, **17**, 523-537.

Lorch, R. F., Jr. (1981). Effects of relation strength and semantic overlap on retrieval and comparison processes during sentence verification. *Journal of Verbal Learning & Verbal Behavior*, **20**, 593-610.

Lorch, R. F., Jr., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 149-157.

Malt, B. C., & Smith, E. E. (1982). The role of familiarity in determining typicality. *Memory & Cognition*, **10**, 69-75.

McCloskey, M. [E.] (1980). The stimulus familiarity problem in semantic memory research. *Journal of Verbal Learning & Verbal Behavior*, **19**, 485-502.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, **6**, 462-472.

McCloskey, M. [E.], & Glucksberg, S. (1979). Decision processes in verifying category membership statements: Implications for models of semantic memory. *Cognitive Psychology*, **11**, 1-37.

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

Mervis, C. B., Catlin, J., & Rosch, E. (1976). Relationships among goodness-of-example, category norms, and word frequency. *Bulletin of the Psychonomic Society*, **7**, 283-284.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 54-65.

Pinker S., & Prince A. (1988). On language and connectionism: Analysis of a parallel distributed model of language acquisition. *Cognition*, **28**, 73-193.

Rips, L. J., Smith, E. E., & Shoben, E. J. (1975). Set theoretic and network models reconsidered: A comment on Hollan's "Features and semantic memory." *Psychological Review*, **82**, 156-157.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, **104**, 192-232.

Rumelhart, D. E., & McClelland, J. L. (1986). PDP models and general issues in cognitive science. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations* (pp. 110-146). Cambridge, MA: MIT Press.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception & Performance*, **3**, 1-17.

Schaeffer, B., & Wallace, R. (1970). The comparison of word meanings. *Journal of Experimental Psychology*, **86**, 144-152.

Shelton, J. R., & Martin, R. C. (1992). How semantic is automatic semantic priming? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 1191-1210.

Smith, E. E. (1978). Theories of semantic memory. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes* (pp. 1-56). Hillsdale, NJ: Erlbaum.

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

Smolensky, P. (1987). The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. *Southern Journal of Philosophy*, **26** (Suppl.), 137-161.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral & Brain Sciences*, **11**, 1-74.

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, **185**, 1124-1131.

Wilkins, A. T. (1970). Conjoint frequency, category size and categorization time. *Journal of Verbal Learning & Verbal Behavior*, **10**, 382-385.

NOTES

1. Use of the notion of "overlap of semantic features" here is for convenience only, to be in keeping with the prevailing semantic theory at the time the models were proposed. It should not be taken as indicating any strong commitment to feature list representations as opposed to other ways of representing semantic content, such as frames or schemas.

2. The data for Experiment 1 were in fact collected within a year or two of the data reported by Hampton and Gardiner (1983).

3. The labeling of a response as an "error" is not always appropriate in tasks where categorization could be a matter of opinion (Hampton, 1979; McCloskey & Glucksberg, 1978).

4. Whether these long latencies were excluded or truncated had a minimal effect on the pattern of results reported below, which, in this case, were based on a total of some 15,000 data points. The same holds true for the results of Experiment 2.

5. Lorch and Myers (1990) pointed out that regression analyses applied to means across subjects are liable to overestimate the significance of independent variables, since they exclude the subject × item interaction variance from the error term. However, their recommended procedure (analyzing the data for each subject separately) runs into the problem of missing values (the relatively high error rates mean that positive RTs would be sampled from different sets of materials for each subject). The analysis of error rates would also not be possible in this case, since it depends on data from the whole group. For technical reasons, the individual RTs were not in any case available for analysis. The present analyses therefore used mean categorization time as the dependent variable, and significance levels should therefore be interpreted with caution. The present study does, however, have the compensatory value that it does not ignore the category × item interaction but allows for the separate analysis of each category. Type 1 errors should appear as a random pattern of significant effects across categories, so, to the extent that a consistent pattern appears, it may be taken as evidence for the validity of the results.

6. By "the wrong sign" is meant that there was a one-tailed prediction made that high-typicality, high-production-frequency, high-familiarity, high-word-frequency, well-known, and unambiguous words would be faster to categorize.

APPENDIX A
Categories and True Items Used in the Item–Category Pairs of Experiment 2

	Typicality		Production Frequency	
Category	High	Low	High	Low
Bird	nightingale	ostrich	eagle	cuckoo
	swift	penguin	hawk	dove
	cuckoo	puffin	duck	peacock
	dove	emu	swallow	turkey
Clothing	jeans	tie	hat	apron
	jacket	gloves	tights	pyjamas
	suit	scarf	socks	bikini
	cardigan	belt	tie	suit
Fish	herring	shark	plaice	pilchard
	sole	eel	eel	piranha
Flower	marigold	dandelion	chrysanthemum	lilac
Food flavoring	ginger	chocolate	thyme	mint
	garlic	thyme	salt	saccharin
Fruit	tangerine	pomegranate	pear	watermelon
	apricot	date	mango	apricot
	mandarin	avocado	peach	mandarin
			pomegranate	satsuma
Furniture	suite	deck-chair	stool	suite
	couch	shelves	sideboard	bench
			cabinet	couch
Insect	cockroach	centipede	spider	locust
	earwig	spider	cockroach	gnat
Sport	basketball	croquet	hockey	pingpong
	baseball	canoeing	riding	snooker
	pingpong	fishing		
	soccer	riding		
Vegetable	leek	pumpkin	turnip	sweetcorn
Vehicle	motorbike	aeroplane	aeroplane	jeep
	van	ship	bus	ambulance
	jeep	tractor	lorry	taxi
	taxi	boat	bicycle	scooter
Weapon	grenade	dart	knife	shotgun
	revolver	whip	spear	machine-gun
	flick-knife	rocket	sword	revolver

(*Continued on next page*)

APPENDIX B
Categories and False Items Used in the Item–Category Pairs of Experiment 2

Category	Related False	Unrelated False
Bird	bat	France
	fly	diesel
Clothing	nylon	symphony
	handbag	bronze
Fish	whale	physics
	lobster	Germany
Flower	nutmeg	ball
Food flavoring	martini	director
	flour	puzzle
Fruit	rhubarb	Paris
	cucumber	trumpet
Furniture	painting	blue
	carseat	cobra
Insect	lizard	cocoa
	snail	coal
Sport	ballet	oxygen
	singing	oboe
Vegetable	almond	corporal
Vehicle	surfboard	Brussels
	missile	zebra
Weapon	forgery	ice cream
	homicide	puppet

Note—Illustrative examples are shown only for false items. Frequency across categories was roughly comparable between true, related false, and unrelated false items.

(Manuscript received October 25, 1995;
revision accepted for publication September 27, 1996.)