Associative and Similarity-based Processes in Categorization Decisions

James A. Hampton

City University, London

Address for correspondence:

Dr James A. Hampton,

Psychology Department,

City University,

Northampton Square,

London, EC1V OHB, England.

To appear in *Memory & Cognition 1997*

Abstract

Two experiments were directed at distinguishing associative and similarity-based accounts of systematic differences in categorization time for different category 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, while under conditions of easy discrimination prior exposure to the category exemplars just affected 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.

Associative and Similarity-based Processes in Categorization Decisions

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 atypical words (Hampton, 1979; Smith, Shoben & Rips, 1974). Typical instances also tend to be items which are 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 proposes that "good" category members are those that have been most often associated with the category in the past. For example, non-analytic 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) termed 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 instance with the category concept (see Chang, 1986; 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 characteristic feature model, or McCloskey & 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 items are categorized more rapidly than atypical items since evidence for a positive decision accrues more rapidly for these items. According to a "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 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 & 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 & Shaffer, 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 by 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, Catlin & Rosch, 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 item-category associative links, and hence the ease with which the items can be retrieved as members. PF reflects the associative aspect of category structure which (following traditional associationist theories) would correspond to frequency of cooccurrence 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 which are commonly encountered but are dissimilar from the prototype, or alternatively others which 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 items with a higher category production frequency (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 participants produced false completions to category sentences. Alternatively Casey (1992), Hampton (1979), Larochelle and Pineau (1994), and Smith, Shoben and Rips (1974) found similar effects on categorization time for high versus low typicality items. The more typical an item is in a category, then the faster is a positive categorization -- and for false category statements, the more similar a non-member 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 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 the 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 items 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 only played a role in the Instance-Category presentation order, where participants would have been more able to guess the true category before it appeared when the item-category pair had high Category Dominance. Larochelle and Pineau (1994) carefully review methodological differences amongst the different studies which could explain the discrepancy in results. Amongst 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 4 and 8 items 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 & Pineau (1994) 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 strategic guessing effects or the retrieval of earlier decisions which 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 9 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 a study 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 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 participant population as used in the present study. [2] These norms provide a large and representative sample of the available category members in each of twelve 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 non-instances presented one at a time in a blocked fashion. This procedure reduces the likelihood that participants are trying to guess the stimuli in advance, (the chance of a correct guess would be about 1%), and also 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 inter-item 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 (a) association-based retrieval of prestored "is a" relations and (b) 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 item 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 items in the twelve 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

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

Design. The twelve categories used by Hampton and Gardiner (1983) were divided into two sets of six, minimizing the apparent similarity between categories within each set. Each participant categorized lists of words for one of the sets. Each list was presented as a block with items randomized for each participant within blocks, and the order of lists was balanced across subjects. Mean response times 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, 3 groups of participants were employed. One group were given twelve category names, and had to generate as many examples of each category as they could in a fixed time. Production frequency was based on this group. A second group rated a list of between 37 and 55 category members for each category (sampled independently of the category exemplar production task), for typicality on a six point scale. A third group rated the same lists of items organized in the same categories for familiarity on a six point scale. Instructions pointed out the difference between the dimensions of typicality, familiarity and frequency of occurrence in order to help participants to focus attention on the relevant dimension. Although not part of "standard" typicality or familiarity instructions, this aspect of the Hampton and Gardiner (1983) 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, participants 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 current 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, 3 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 participants perform a clustering-by-similarity task on the 56 categories used by Battig and Montague (1969). For example, for the category Clothing, the related false items were from the category of Footwear, while for Food Flavourings 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 3 false categories. Since the number of positive items varied between categories, the final lists contained between 68 and 110 words.

<u>Procedure</u>. Participants 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

participant. The participant pressed one of two response keys as rapidly as possible, to indicate whether each item belonged in the named category or not. After completing each list, participants were given a two minute rest. Instructions emphasized the importance of making as few errors as possible. The whole session took about 45 minutes.

Results

To remove the undue effect of extreme response times, 15 latencies (0.1%) of less than 250 ms were excluded from the analysis of mean correct 'Yes' response times, and 33 latencies (0.2%) of over 3000 ms were truncated to 3000 ms. [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 ms for unrelated category items, 795 ms for slightly related items, and 798 ms for strongly related items. Mean true categorization time across all categories was intermediate between these levels at 762 ms, and varied across categories from 696 ms for Birds to 880 ms for Insects. However within each category, mean true categorization time for individual items varied widely, from a low of 600 ms to a high of 2000 ms. 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 and the second set of 15 participants judging each category list. Corrected reliabilities varied from .63 for SPORTS, to .88 for 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 response time -- word frequency taken from Kucera 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.)

INSERT TABLE 1 ABOUT HERE

The two variables of greatest theoretical interest, Typicality and PF, were equally well correlated with categorization time overall, at 0.66 and -0.65 respectively. Familiarity was less well correlated with categorization time (average 0.56), although for 2 categories - CLOTHING and FLOWERS - familiarity had the highest correlation.

INSERT TABLE 2 ABOUT HERE

Table 2 gives the standardized regression coefficients (beta) for the regressions predicting categorization time from seven variables: PF, typicality, familiarity, word frequency and length, as defined previously, together with UNKNOWN which was defined as the number of participants 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 (for example 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 step-wise fashion any variables entered with the wrong sign [6], or with a non-significant regression weight (alpha = .05, one-tailed). (Only 3 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 just 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 squared, 4 categories identified typicality as a significant predictor, and 2 picked out PF.

The same general pattern of weights emerged when all the 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 ms slower), Furniture (84 ms), and Food flavourings (82 ms). Subsequent forward steps then included the following variables (with associated beta 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 attributable 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 ms for typicality, and 103 ms for PF, while each extra percent of participants not knowing an item increased categorization time by 6 ms, and each letter of a word took an extra 7 ms 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 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 participants.

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 = 0.778), and in the correlation across categories ($r= 0.77$, $n=12$, $p<.005$). 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: PF (log transformed), typicality, familiarity, UNKNOWN and AMBIGUITY, all as defined previously. To disconfound the typicality scale from the proportion of participants rejecting a category exemplar, mean typicality values were recalculated from the norms for this analysis by excluding any participants who gave a rating of 6 (= "not in the category"). The mean typicality values thus reflected the mean typicality judgement of those participants who believed that the item was a category member.

Across all categories, typicality was much the best predictor of response rate (beta = 0.538), with UNKNOWN (0.332) and AMBIGUITY (0.129) also predictive. Multiple R was 0.692. 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 people rejecting 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 supported 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 items per category, and a list-wise 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 a 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 categorization 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' (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 which 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 which 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 showed 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 participant 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 item 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 instance-category pairs in a list to be categorized were all unrelated, then true response times were both faster and also less sensitive to differences between typical and atypical category members. When the relatedness of false instance-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 (1979) in fact used PF as the basis for selecting high and low typicality items for their categories so that their items differed in both PF and typicality. For the current study, the strong prediction can be made that their result should be found for materials that differ in typicality, but should <u>not</u> be found for materials that differ only in PF. Experiment 2 also answered a potential criticism of McCloskey & 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 response times, and could therefore have merely reflected the skewed distribution of reaction times in general. Since Experiment 2 predicts the increase will occur specifically for differences in typicality and <u>not</u> 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) found, as predicted, that introducing related false items into a list of true instance-category pairs increased the typicality effect on categorization times.from 18ms to 48ms, but did not increase the PF effect (40ms vs 37ms). Two related false conditions were used -- one in which all non-members were related, and a second in which only half the non-members were related. Both led to an increase in the typicality effect, but in the all-related condition (corresponding the McCloskey & Glucksberg's Experiment 2) there were some participants 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 <u>any</u> 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 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 participants, 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 <u>a priori</u> 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 participants 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 participant would be asked to judge the instance-category pair SWIFT-CREATURE. Later, in the main part of the experiment the participant 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 items should therefore show greater priming than the high PF items, since the latter would already have easily accessible instance-category links. Since the prior exposure did <u>not</u> directly involve categorization of the instance in the target category, it was predicted that the typicality effect would remain unaffected by this priming manipulation. Deciding for example that an atypical item 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 item like 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 2x2 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 participant 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

Participants. Participants 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 4 equal groups of 24 participants each.

Design. The design incorporated two between-group factors. The first was Measure (typicality versus 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 participants. That is, half the participants took part in conditions considering effects of typicality on categorization time, and half 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 participants' decision criterion for making a categorization response. In one set of conditions, participants 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 participants 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, and in the subsequent task, 60% of the false instance-category pairs were indeed related. To summarize, there were four groups of participants 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 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 participants.

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), matched for typicality (1.85 and 1.86) and familiarity (1.45 and 1.53). The pairs were taken from all 12 categories in Hampton and Gardiner (1983), and are listed in the Appendix. Each category always occurred equally often with high and with low items across different conditions. The initial priming session consisted of a categorization task, similar to that used for the main test. Sixteen of the 32 High 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, Manmade-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 which 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 5 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 8 filler items which while true for the priming session would be false pairs for the test session - to discourage participants from using the response made in the priming session as a way of predicting the response in the test. With 10 practice, 32 true (16 high and 16 low), 32 false and 8 fillers, 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 neighbouring 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 the Appendix. 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 the 8 fillers from the priming session which 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 the previous experiment. Participants were given one of two written sheets of instructions, according to whether they were in the Speed or 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 participant was not aware of this, but 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 upper case, with the instance displayed simultaneously with, and directly above, the category name. A warning asterisk signalled 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 subjects preferred hand.. After 41 trials, participants 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 half way through the session. After the experiment, participants were debriefed, and asked to tell of any ambiguous items or other problems they may have encountered.

Results

Latencies less than 250 ms were excluded, and latencies over 3000 ms (less than 1%) were truncated to 3000 ms. (An alternative analysis was run in which long latencies were excluded if greater than 3 standard deviations above the mean calculated separately for each participant, with the same general results). Table 3 shows the full set of mean categorization times for each condition in the Experiment.

INSERT TABLE 3 ABOUT HERE

A 5-way ANOVA was conducted of the complete design, with factors of Centrality (high versus low values of either PF or Typicality), measure (PF versus Typicality), priming (primed versus unprimed), criterion (Speed versus 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, and participants 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 (a)priming, and (b)criterion, 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 ms. 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, shown in the upper half of the table, the effect was exactly as predicted. Under Speed conditions, the typicality effect was 22 ms (primed) and 23 ms (unprimed). Under Accuracy conditions it rose to 100 ms (primed) and 114 ms (unprimed). A 4-way 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 participants 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 ms for Speed conditions and 57 ms for Accuracy conditions. Thus while the typicality effect for unprimed pairs increased by some 91 ms as criterion was manipulated, the PF effect increased by only 15 ms. 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 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 participant groups. The manipulation was therefore clearly effective, speeding categorization time by some 35 to 105 ms. 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 (54 ms and 53 ms) and for low typicality pairs (55 ms and 67 ms). In the PF-Accuracy condition, priming was greater but again did not differentiate between high PF (102 ms) and low PF (105 ms) pairs. For the PF-Speed condition however there was an interaction between priming and PF. Priming was stronger for low PF (76 ms) than for high PF (35 ms) pairs. Put another way, the PF effect (42 ms) observed for unprimed pairs was completely removed (1 ms) when pairs were primed in the Speed condition. A 3-way ANOVA of the PF-Speed condition with priming, word-set, and PF as factors confirmed a significant interaction between priming and PF with participants as random factor ($F(1,22)=5.17$, $MSe=1893$, $p<.05$), and with words as random factor ($F(1,60)=4.03$, $MSe=2345$, $p<.05$).

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 (76 ms), intermediate for the high and low typicality items which had medium to low PF (54 and 55 ms), and least for high PF items (35 ms). The priming factor therefore dissociated PF and typicality as measures, in that typicality showed zero interaction with priming, under either Speed or Accuracy conditions, whereas PF showed the prediction 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 ms) for both high and low PF items. This interaction between the two major manipulations of the experiment was unexpected. Given that participants were responding more cautiously in general in the Accuracy condition, the increased size of the priming effect may be the result of the need to access a greater amount of relevant semantic information of all kinds (including property as well as 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 items 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 Accuracy condition, priming helped both high and low PF items equally.

INSERT TABLE 4 ABOUT HERE

False items. Response times for correctly rejecting false items are shown in Table 4. Since the same false item materials were used in both the typicality and PF comparison conditions, they have been averaged in the third and sixth lines of the table. First, as expected, Unrelated False items (1023 and 1076 ms) were faster than Related False items (1314 and 1365 ms) in the Accuracy condition. These same Unrelated False items were much faster again (765 and 807 ms) 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 between 42 and 53 ms. As this priming effect is unaffected by the relatedness of the false items (compare effects of 51 ms for related false and 53 ms 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 5-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 participants. These were Centrality, (across subjects $F(1,88) = 8.88$, $p < .005$, across words $F(1,120) = 4.52$, $p < .05$), and the 3-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 make 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 items (mean typicality = 2.93) intermediate for the PF materials (typicality = 1.85) and least for high typicality items (typicality = 1.42). By contrast in the Speed condition error rates were highest for the low PF items (mean PF = 4.5), slightly lower for the typicality materials (PF = 13.6) and least for high PF items (PF = 33.6).

Discussion

The results of Experiment 2 can be summarized as follows. First, changing the relatedness of false items and encouraging participants to raise their decision criterion, had the effect of slowing down atypical category members more than typical members. High fxand 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 reaction time, discovered by McCloskey and Glucksberg (1979), works mainly by slowing down atypical items, as opposed to low PF items. 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 items, fewest made to typical items, 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 items. With the emphasis placed on speed, and no related distractors, participants 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 instance and category may have been used as the basis to erroneously reject a low PF item from the category.

The new factor introduced in Experiment 2 was the repetition priming manipulation. Predictions for this manipulation could only be based 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 which 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, primimg with an earlier decision that the item was in a more superordinate category was completely ineffective in changing the difference in response time between high and low typical items (that is both sets of items were equally primed by the repetition). For PF effects, the priming was effective in removing the difference between the categorization time for high and low PF items, but this only occurred in the low criterion Speed condition. This result is consistent with the suggested strategy for this condition, that participants 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 items 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 items, and a priming effect that was greater for low PF than for high PF items. In the Accuracy condition however, this strategy would have lead to unacceptably high error rates as 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 <u>featural</u> information needed for a category decision based on similarity. Only the high PF items, which are more likely to have specific "is a" links to the category, would then be categorized on the basis of specific category associations, while 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 response time, while 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 items 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 which 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 items was magnified by the increased difficulty of the task, as was the difference in error rates between the two kinds of items. By contrast low PF items did not show either of these effects relative to high PF items in the Unprimed condition, and in fact were less prone to error when related false items were included. The experiment thus clarified McCloskey & 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 born out, but only in the Speed Condition. Priming with a superordinate categorization removed the difference in categorization time for high PF and low PF items 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 items in the Accuracy condition, but more errors made to low PF items in the Speed condition. The occurrence of errors to low PF items 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 current 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 items 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 categorizion 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 reaction time and error rates seen whenever there are false related items in the experiment.

The direction taken by this research has been towards 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 towards this goal, indicating the need for independent consideration of association and similarity based effects within the framework of modelling 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 associative 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, participants 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, participants may simply decide category membership 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 participants 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, and 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 and Performance, 5, 252-259.

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

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 and 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 and Cognition, 12, 257-267.

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

Collins, A.M., & Quillian, M.R. (1969). Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior, 8, 240-248.

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 and Cognition, 6, 622-628.

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

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

Hampton, J.A. (1988). Evidence for a two process model of categorization. Paper presented to the 29th Annual Convention of the Psychonomic Society, Chicago IL, November.

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. 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., and Smith, E.E. (1975). Cross-category structure in semantic memory. Memory and Cognition, 3, 591-594.

Hollan, J.D. (1975). Features and semantic memory: Set theoretic or network models? Psychological Review, 82, 154-155.

Kucera, H., & Francis, W.N. (1967). Computational analysis of present-day American English. Providence, R.I.: Brown University Press.

Larochelle, S.,& Pineau, H. (1994). Determinants of response times in the semantic verification task. Journal of Memory and 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 and 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 and Verbal Behavior, 20, 593-610.

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

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

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

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

McCloskey, M., & 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 and 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 and the PDP research group (Eds.), Parallel Distributed Processing: Explorations in the microstructure of cognition. Vol. 1: Foundations. 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 and 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 and Cognition, 18, 1191-1210.

Smith, E.E. (1978). Theories of semantic memory. In W.K. Estes (Ed.) Handbook of Learning and Cognitive Processes (Vol. 5.) Hillsdale, N.J.: Lawrence 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. The Southern Journal of Philosophy, Supplement, 26, 137-161.

Smolensky P. (1988). On the proper treatment of connectionism. Behavioral and 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 and Verbal Behavior, 10, 382-385.

Appendix

Item-Category Pairs Used in Experiment 2

<u>True items</u>

TYPICALITY CONDITION

Category	High Typicality		Low Typicality	
BIRD	Nightingale	Cuckoo	Ostrich	Puffin
	Swift	Dove	Penguin	Emu
CLOTHING	Jeans	Suit	Tie	Scarf
	Jacket	Cardigan	Gloves	Belt
FISH	Herring	Sole	Shark	Eel
FLOWER	Marigold	-	Dandelion	-
FOOD-FLAVOURING	Ginger	Garlic	Chocolate	Thyme
FRUIT	Tangerine	Mandarin	Pomegranate	Avocado
	Apricot	-	Date	-
FURNITURE	Suite	Couch	Deck-chair	Shelves
INSECT	Cockroach	Earwig	Centipede	Spider
SPORT	Basketball	Pingpong	Croquet	Fishing
	Baseball	Soccer	Canoeing	Riding
VEGETABLE	Leek	-	Pumpkin	-
VEHICLE	Motorbike	Jeep	Aeroplane	Tractor
	Van	Taxi	Ship	Boat
WEAPON	Grenade	Flick-knife	Dart	Rocket
	Revolver	-	Whip	-

Appendix (continued)

PRODUCTION FREQUENCY CONDITION

Category	High PF		Low PF	
BIRD	Eagle	Duck	Cuckoo	Peacock
	Hawk	Swallow	Dove	Turkey
CLOTHING	Hat	Socks	Apron	Bikini
	Tights	Tie	Pyjamas	Suit
FISH	Plaice	Eel	Pilchard	Piranha
FLOWER	Chrysanthemum	-	Lilac	-
FOOD-FLAVORING	Thyme	Salt	Mint	Saccharin
FRUIT	Pear	Peach	Watermelon	Mandarin
	Mango	Pomegranate	Apricot	Satsuma
FURNITURE	Stool	Cabinet	Suite	Couch
	Sideboard	-	Bench	-
INSECT	Spider	Cockroach	Locust	Gnat
SPORT	Hockey	Riding	Pingpong	Snooker
VEGETABLE	Turnip	-	Sweetcorn	-
VEHICLE	Aeroplane	Lorry	Jeep	Taxi
	Bus	Bicycle	Ambulance	Scooter
WEAPON	Knife	Sword	Shotgun	Revolver
	Spear	-	Machine-gun	-

Appendix (continued)

False items (examples)

Category	Related False		Unrelated False	
BIRD	Bat	Fly	France	Diesel
CLOTHING	Nylon	Handbag	Symphony	Bronze
FISH	Whale	Lobster	Physics	Germany
FLOWER	Nutmeg		Ball	
FOOD-FLAVORING	Martini	Flour	Director	Puzzle
FRUIT	Rhubarb	Cucumber	Paris	Trumpet
FURNITURE	Painting	Carseat	Blue	Cobra
INSECT	Lizard	Snail	Cocoa	Coal
SPORT	Ballet	Singing	Oxygen	Oboe
VEGETABLE	Almond		Corporal	
VEHICLE	Surfboard	Missile	Brussels	Zebra
WEAPON	Forgery	Homicide	Ice-cream	Puppet

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

Author Notes

Thanks are due to 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 to 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.

Footnotes

[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 & Gardiner (1983).

[3]. The labelling 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, as they exclude the subject x item interaction variance from the error term. However, their recommended procedure (analysing the data for each subject separately) runs into the problem of missing values (the relatively high error rates mean that positive reaction times 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 response times were not in any case available for analysis. The present analyses therefore used mean CT 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 x item interaction, but allows for the separate analysis of each category. Type I 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.

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 FLAVOURINGS	-.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
Mean	-.65	.66	.56	-.15	.17

(Note: PF = Log Production Frequency, TYP = Typicality, FAM = Familiarity, WF = Word Frequency, LEN = Word length)

Table 2

Standardized Regression Weights (beta) for each of the Significant Predictors of Categorization Time in Expt 1, and R for the Optimal Regression Equation.

	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 FLAVOURINGS	_	-.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--Production Frequency, FAM--Familiarity, WF--Word Frequency, LEN--Word Length, UNK--Unknown, AMB--Ambiguity.)

Table 3

Mean and (Standard Deviation) Categorization Times, and Percentage Error Rates, for True Items in each Condition in Experiment 2.

Centrality Measure	Instruction and Priming Conditions					
	Speed			Accuracy		
	Primed	Unprimed	Priming Effect	Primed	Unprimed	Priming Effect
Typicality						
High Typ	754 (197)	808 (217)	54	961 (231)	1014 (231)	53
% Error	4	6		3	3	
Low Typ	776 (212)	831 (223)	55	1061 (290)	1128 (327)	67
%Error	4	6		7	8	
Typ Effect	22	23		100	114	
PF						
High PF	695 (104)	730 (106)	35	925 (224)	1027 (307)	102
% Error	3	4		4	4	
Low PF	696 (95)	772 (108)	76	979 (269)	1084 (304)	105
% Error	5	7		4	4	
PF Effect	1	42		54	57	

Table 4

Mean and (Standard Deviation) for Reaction Times for False Items in Experiment 2.

Response times	Speed			Accuracy		
	Primed	Unprimed	Priming Effect	Primed	Unprimed	Priming Effect
Related False						
Typicality condition	-	-		1336 (319)	1402 (369)	66
PF condition	-	-		1292 (391)	1328 (382)	36
Mean	-	-	-	1314	1365	51
Unrelated False						
Typicality condition	788 (191)	830 (211)		1025 (256)	1054 (269)	29
PF condition	743 (110)	784 (136)		1020 (282)	1099 (343)	79
Mean	765	807	42	1023	1076	53