*Bulletin of the Psychonomic Society*
*1984, 22 (4), 324-326*

# Reliability of retrieval from semantic memory: Common categories

FRANCIS S. BELLEZZA
*Ohio University, Athens, Ohio*

**Permanent verbal knowledge about objects in the physical world and about the meaning of words resides in semantic memory. Little research has been done to determine how reliably such information can be retrieved from semantic memory. In the present experiment, an estimate of reliability was made. The method used was to ask subjects to perform the same retrieval task in each of two sessions separated by 1 week. In each session, the subjects were asked to generate instances of common categories. The mean correlation between the contents of the two recalls was found to be .69. As expected, consistency in the content of recall from session to session was greater within subjects than between subjects, for which the correlation value was .44. The results seem to indicate that retrieval of information from semantic memory is a probabilistic process that occurs with only a modest amount of reliability.**

The distinction between episodic and semantic memory suggests that information in semantic memory is relatively stable and accessible (Tulving, 1972). This particular assumption, however, has not been widely tested. The purpose of the present experiment was to determine how reliably information can be retrieved from the semantic memory of individuals with regard to common categories. The method used was that of information generation. Subjects provided responses to common category labels by writing down instances of the categories presented. One week later they were tested on the same task using the same labels. The written responses were then compared with regard to their content. The overlap in content, as measured by the common-element correlation, indicated how reliably the same information about a particular category could be retrieved from semantic memory.

## METHOD

### Subjects

Thirty-four undergraduates enrolled in introductory psychology courses at Ohio University volunteered for extra course credit.

### Procedure

All subjects were tested twice as a group, with Session 2 following Session 1 by 1 week. The procedure was the same in both sessions. On the top of each page in the test booklets was printed the name of a category taken from Battig and Montague (1969). The 12 categories used typically were made up of concrete objects and are listed in Table 1. In each session, the subjects were given 3 min to write down as many instances of each category as they could in the order that the instances came to mind. The presentation order of the category labels was random in each session.

Portions of this research were presented at the meeting of the Psychonomic Society in Minneapolis, Minnesota, in November 1982. This research was supported in part by a grant from the Field-Wiltsie Foundation. Thanks go to Kathy Kamin, Sylvia Sims, Rosalind Whatley, and Cathy Young for their assistance in collecting and scoring the data. Thanks also go to Ohio University Computer and Learning Services for making computer time and their facilities available. Requests for reprints should be sent to Francis S. Bellezza, Department of Psychology, Ohio University, Athens, OH 45701.

## RESULTS

A 2 x 12 analysis of variance was performed on the mean number of instances recalled for each category, with the first factor being session and the second factor representing category label. Both factors were within-subjects factors. Slightly more instances were recalled in Session 2 (mean of 21.55) than in Session 1 (mean of 20.73) [$F(1,33) = 4.26$, $p < .05$, $MSe = 32.65$]. Also, there were significant differences among the mean number of instances recalled for the various category labels [$F(11,363) = 137.02$, $p < .001$, $MSe = 33.16$]. The mean number of instances recalled for each category is shown in Table 1. Tukey's HSD test (Kirk, 1968) indicated that a difference of 3.23 was needed for any two means to be significantly different. There was a significant session x category interaction [$F(11,363) = 5.13$, $p < .001$, $MSe = 8.29$]. Posttests using Cicchetti's (1972) procedure showed that the subjects recalled more instances of body parts and birds in Session 2 than in Session 1. These two categories were the first two categories tested during Session 1, so the lower numbers of instances generated may indicate that the subjects became more proficient at the generation task after practice on the first two category labels.

### Intrasubject Reliability

To determine the degree of overlap in the category instances recalled by any one subject during the two sessions, a common-element correlation was used (Deese,

Table 1
Relations Between Recall in Session 1 and in Session 2 for the Category Labels Presented

| Category | Mean Instances Recalled | Within-Subjects Overlap | P(Recall 2\| Recall 1)* | Correlation Between Recall 1 Order and Recall 2 Order | Probability of Recalling Adjacent Items | Between-Subjects Overlap |
|---|---|---|---|---|---|---|
| Animals | 22.4 | .68 | .84 | .52 | .14 | .50 |
| Birds | 16.5 | .72 | .87 | .57 | .12 | .38 |
| Body parts | 32.3 | .69 | .85 | .40 | .10 | .51 |
| Clothing | 23.7 | .67 | .81 | .33 | .10 | .46 |
| Flowers | 12.0 | .71 | .83 | .42 | .10 | .44 |
| Fruits | 17.8 | .79 | .92 | .59 | .15 | .60 |
| Furniture | 17.7 | .66 | .82 | .72 | .08 | .37 |
| Girls' names | 40.5 | .50 | .65 | .60 | .08 | .18 |
| Musical instruments | 19.2 | .80 | .91 | .46 | .10 | .57 |
| Sports | 22.7 | .71 | .88 | .57 | .10 | .48 |
| Trees | 13.0 | .69 | .81 | .54 | .09 | .42 |
| Vehicles | 15.9 | .65 | .77 | .38 | .12 | .32 |

**This value represents the estimated probability of recalling an item in Session 2 given that it was one of the first five items recalled in Session 1.*

1965; McNemar, 1969, pp. 145-146). To compute this value for each category for each subject, the number of category instances recalled in both sessions was divided by the square root of the product of the total number of category instances recalled in Session 1 and the total number of instances recalled in Session 2 (the geometric mean). If the number of instances recalled in Session 1 is identical to the number recalled in Session 2, then the common-element correlation represents the proportion of the instances recalled that are common to both sessions. This measure of correlation varies between the values of 0 and 1. Instances were considered the same regardless of spelling errors or differences in singular versus plural form.

The mean overlap score over all subjects and categories was .69. An analysis of variance on the 12 overlap means showed category to be a significant factor [$F(11,363) = 18.12$, $p < .001$, $MSe = .011$]. The mean score for each category is shown in Table 1. A Tukey HSD test indicated that a difference of .081 was necessary for two overlap means to be significantly different.

The overlap values indicate that there is a moderate amount of correspondence between retrieval of category instances in the two recall sessions. Mean overlap scores ranged in value from .50 to .80. An additional analysis was performed to determine if the instances generated early in Session 1 rather than late in Session 1 were more likely to be generated during Session 2. The probability of recalling an instance any time during Session 2 was computed given that it was one of the first five instances generated in Session 1. This was done for each category by combining the data of the 34 subjects. These estimated probabilities are shown in Table 1. The mean probability for all 12 categories combined was .83. If the same procedure is used to estimate the probabilities of recall for words recalled in other serial positions in Session 1, the following values are found: for Positions 6 to 10, .73; for Positions 11 to 15, .66; for Positions 16 to 20, .63; and for Positions 21 to 25, .50. Not enough recall data was available to estimate probabilities of recall given that a word was recalled in Session 1 beyond Serial Position 25.

It appears from Table 1 that as the number of instances generated for a category increases, then the overlap score decreases. To determine if this was the case, a correlation was computed between the 12 category sizes and the 12 category overlap scores from the data of each subject. Each one of these 34 correlations was then transformed using Fisher's z transformation (McNemar, 1969). An analysis of variance showed the mean correlation value of $-.39$ to be significantly less than zero [$F(1,33) = 60.66$, $p < .001$, $MSe = .093$]. This negative correlation would occur if the subjects were randomly drawing category instances from a population of instances for each category.

**Organization of Instance Recall**

Although the subjects were not recalling precisely the same category instances in the two sessions, the question remains as to what similarity, if any, there was in the organization of the instances that were recalled in both sessions. For example, were the category instances recalled in both sessions recalled in the same relative order? To test this hypothesis, the responses common to both tests of recall were ranked by order of output for Session 1 and for Session 2 for each subject-category combination. These ranks were correlated using Spearman's rank-correlation coefficient (McNemar, 1969). The mean correlation was .51, which was significantly greater than zero [$F(1,330) = 158.98$, $p < .001$, $MSe = .239$]. However, there were no significant differences among the categories [$F(11,338) = 1.65$]. The mean output-output correlation for each category label is shown in Table 1.

The output-output correlations indicate that the relative orders of common responses in the two sessions were similar. An additional analysis was performed to determine the probability that any pair of adjacent responses made in Session 1 was also made in the same order in Session 2. The overall mean probability of this

occurring was found to be .11, which was small but significantly greater than zero [$F(1,33) = 306.58$, $p < .001$, $MSe = .014$]. An analysis of variance showed that there were also significant differences in this probability value among the categories [$F(11,330) = 2.41$, $p < .005$, $MSe = .007$]. The value for each category is shown in Table 1. A Tukey HSD test indicated that a difference of .066 was needed for two values to be considered significantly different.

### Intersubject Reliability

The measures presented so far reflect the degree to which one particular person retrieves the same category instances in two recall sessions separated by a week. Also of interest, however, is the degree to which different subjects recall the same category instances. To obtain a measure of intersubject agreement, the subjects were randomly paired using the data from Session 1 and were again randomly paired using the data from Session 2. The overlap scores for the two individuals in each pair were then computed for each category. A 2 x 12 analysis of variance was then performed on the means, with the first factor being session and the second factor being category. Session was a between-pairs factor, and category was a within-pairs factor. Only category was a significant source of variation [$F(11,352) = 31.98$, $p < .001$, $MSe = .014$]. The between-subjects overlap means are shown in Table 1. The overall mean overlap score was .44, which was significantly greater than zero [$F(1,30) = 2,019$, $p < .001$, $MSe = .039$]. A Tukey HSD test indicated that two overlap scores were significantly different if the difference was greater than .092.

## DISCUSSION

The results indicate that instances of categories are stored in semantic memory such that they are not retrieved in a highly systematic manner. Approximately 69% of the category instances recalled in one session were recalled in the other. One might argue that the subjects first retrieved instances that were clearly typical category members and then retrieved instances that were more atypical or ambiguous as to category membership (McCloskey & Glucksberg, 1978, 1979; Oden, 1977; Rosch, 1975; Rosch & Mervis, 1975). However, of the first five instances recalled in Session 1, only 83% were also recalled in Session 2. The modest overlap scores may be the result not only of variability of typicality, but also of the manner in which even the most typical category instances are stored and retrieved from memory.

If a large number of instances was generated for a category, then it was unlikely that an instance was recalled in both sessions. This is evidenced by the −.39 correlation found between category size and overlap score. However, category instances do not seem to be randomly sampled. The mean rank correlation between the retrieval order of instances recalled in both sessions was .51. Also, category instances retrieved in Positions 1 to 5 in Session 1 were more likely to be retrieved in Session 2 than category instances retrieved in Positions 21 to 25. This organization seems to result more from the recall of sets of category instances than from the recall of particular series of category instances. The probability that a particular pair AB of instances retrieved in Session 1 would also be retrieved adjacently and in the order AB in Session 2 was .11.

There exists greater reliability in category recall within subjects, for which the mean overlap score between sessions was .69, than between subjects, for which the mean overlap score between subjects was .44. This result is not surprising and shows that the components that a semantic category comprises vary among individuals.

## REFERENCES


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 Monographs*, **80**(3, Pt. 2).

Cicchetti, D. V. (1972). Extension of multiple-range tests to interaction tables in the analysis of variance: A rapid approximate solution. *Psychological Bulletin*, **77**, 405-408.

Deese, J. (1965). *The structure of associations in language and thought*. Baltimore: Johns Hopkins Press.

Kirk, R. E. (1968). *Experimental design: Procedures for the behavioral sciences*. Belmont, CA: Wadsworth.

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 verbal category membership statements: Implications for models of semantic memory. *Cognitive Psychology*, **11**, 1-37.

McNemar, A. (1969). *Psychological statistics* (4th ed.). New York: Wiley.

Oden, G. C. (1977). Fuzziness in semantic memory: Choosing exemplars of subjective categories. *Memory & Cognition*, **5**, 198-204.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, **104**, 192-233.

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

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory*. New York: Academic Press.


(Manuscript received for publication April 4, 1984.)