Memory & Cognition
1976, 4 (5), 655-660

Further examinations of the category-recall function

ROBERT M. SCHWARTZ
Rio Hondo College, Whittier, California 90608

and

MICHAEL S. HUMPHREYS
Northwestern University, Evanston, Illinois 60201

Subjects were assigned to use either three or six categories and were given either 5 or 10 trials in a word-sorting task. Subsequent to sorting, they recalled as many words as they could. A measure of sorting consistency showed that the six-category sort was a more difficult task than was the three-category sort. Number of categories did not have a significant effect on recall performance regardless of whether 5 or 10 sorting trials were used. A correlational analysis raised questions about the relationship of the sorting tasks to recall performance and clustering.

Mandler (1967) reported a number of experiments from a task in which he observed a linear relationship between the number of categories subjects use in word sorting and the number of words they recall. Generally, in this task the subjects sort words into categories; they are usually allowed to choose the number of categories into which they sort, with the restriction that they use no fewer than two and no more than seven categories. The sorting process is continued until stable categorization is reached, and then the subjects are asked to recall as many words as they can. Mandler (1967) has found that the relationship between number of categories and recall is typically characterized by a slope value of 5 ± 2, which indicates that subjects recall an additional 5 ± 2 words per category used.

There are two possible artifacts in the Mandler (1967) paradigm. First, allowing subjects to choose the number of categories may result in an artifact of self-selection. Second, as sorting is self-paced and continued to a criterion, the number of sorting trials and the time per sorting trial are not experimentally controlled. Schwartz and Humphreys (1972) found limited evidence for the first of these two artifacts. They first gave subjects five sorting trials in which the words were presented at a fixed rate. Subjects were allowed to sort into any number of categories between two and seven and were then asked to recall the words. Then, the subjects were given four sorting trials, again at a fixed rate, on a second list. Here, they were randomly assigned to groups which were required to sort into either three or six categories. On the basis of their List 1 sorts, subjects were classified as high categorizers (they had sorted into five or more categories) or as low categorizers (they had sorted into four or fewer categories). In two experiments high categorizers recalled more on List 2, where choice of category was randomly assigned, than did low categorizers, though this difference was not significant. However, when Melkman (1975) repeated this experiment using a self-paced sorting task and removing the restriction that the List 1 sort be between two and seven categories, there was clear evidence for subject selection. The more categories the subjects elected to use on List 1, the better was their recall performance on List 2.

Mandler (1967) examined the second artifact by computing correlations between amount recalled and both number of trials and time taken to reach a criterion of sorting consistency. Since these correlations were essentially zero, Mandler concluded that number of trials and total time were not responsible for the category-recall functions. The problem with this conclusion is that there are two components to the correlations Mandler calculated: (1) the correlation between recall and trials (time) for those subjects at each level of the number of categories variable, and (2) the correlation between mean recall and mean trials (time) across the levels of the number of categories variable. It seems possible that, within each level of the number of categories variable, the correlation might be negative, as those subjects who find the task easier or who try harder may learn to categorize quicker and recall more. However, if sorting into a large number of categories is a more difficult task than sorting into a small number of categories, the subjects who choose to sort into a large number of categories might take more time and learn more because of this extra time, producing a posi-

The data were collected while the authors were at the University of British Columbia. The data collection was supported by Grant APA 337 from the National Research Council of Canada. Requests for reprints should be sent to R. M. Schwartz, Social Sciences Department, Rio Hondo College, Whittier, California 90608.

Table 1
Example Matrix for Determination of Consistency Measure

	Trial 5: Category			
Trial 4	1	2	3	Total
Category 1	21	0	0	21
Category 2	0	15	9	24
Category 3	0	0	7	7
Total	21	15	16	52

tive correlation across categories. When these two correlations differ in sign, the total correlation is indeterminate, and Mandler's (1967) finding of near zero correlations is not interpretable.

The primary purpose of this study was to see if it is more difficult to sort into six categories than into three. To determine sorting difficulty, a measure of sorting consistency (how similar were the sorts of a given subject from one trial to the next) was devised. A second purpose was to see if the relatively low level of consistency in the Schwartz and Humphreys (1972) experiments was responsible for their failure to find a significant category-recall function when time per trial, number of trials, and category assignment were controlled. Number of sorting trials (5 vs. 10) was used to manipulate the terminal level of consistency achieved by the subjects. The final purpose was to examine the relationship between the three variables of amount recalled, sorting consistency, and amount of clustering (the extent to which the order of recall conforms to the categories the subjects used on the sorting task).

SORTING-CONSISTENCY MEASURE

For a measure of sorting consistency, Mandler (1967) used the percentage of items sorted into the same category on successive trials. This measure, however, has two difficulties. First, especially on early trials when the subject has used many categories, it is sometimes difficult for the experimenter to determine which category on trial n is which category on trial n + 1. Second, the percentage of items sorted identically does not take into account the probability that, by chance, the subject will sort an item into the same category on successive trials. This probabiltiy should vary as a function of the number of categories used and the lengths of the categories used.

The basic unit for the present measure involves pairs of items. Four observed and expected numbers of pairs of items should be obtained: (1) the number of specific pairs of items contained both in the cluster structure on trial n and in the cluster structure on trial n + 1; (2) the number of specific pairs of items contained in the cluster structure on trial n but not contained in the cluster structure on trial n + 1; (3) the number of specific pairs of items contained in the cluster structure on trial n + 1 but not contained in the cluster structure on trial n; and (4) the number of specific pairs of items contained neither in the cluster structure on trial n nor in the cluster structure on trial n + 1. Once the observed and expected values have been determined, a statistic similar to chi-square with 1 df may be obtained. The pairs entering into the calculations are not independent, so the statistic is not distributed exactly as chi-square with 1 df.

To illustrate the determination of the observed and expected values, a matrix derived from typical sorting protocols is contained in Table 1. The matrix was derived from Trials 4 and 5 of the sorts of a subject assigned to use three categories in the present experiment. The row totals represent the number of words in each of his three categories on Trial 4, and the column totals represent the number of words in each of his three categories on Trial 5. Denote the row totals as A_i and the column totals as A_j. Both ϵA_i and ϵA_j should equal the number of items sorted, which in the example is 52. The total number of pairs contained in the Trial 4 cluster structure is $\epsilon\binom{A}{2}i)$, and the total number of pairs contained in the Trial 5 cluster structure is $\epsilon\binom{A}{2}j)$. In the example there are 507 pairs of items contained in the Trial 4 cluster structure and 435 pairs of items contained in the Trial 5 cluster structure.

The cells, denoted A_{ij}, represent the number of items contained in category i on Trial 4 and category j on Trial 5. The observed number of pairs of items common to the two cluster structures is $\epsilon\binom{A}{2}ij)$, which in the example is 372.

Determination of the expected number of pairs common to the two cluster structures involves the total number of possible pairs $\binom{N}{2}$. In the example there are 1,326 possible pairs. The expected number of pairs common to the two cluster structures is $[\epsilon\binom{A}{2}i)\epsilon\binom{A}{2}j)]/\binom{N}{2}$, which in the example is 166.32. Constraints imposed by the total number of pairs contained in the Trial 4 cluster structure, the total number of pairs contained in the Trial 5 cluster structure, and the total number of possible pairs determine the observed and expected values for the number of pairs unique to the Trial 4 cluster structure, the number of pairs unique to the Trial 5 cluster structure, and the number of pairs contained in neither cluster structure. The four observed and expected numbers of pairs necessary for calculation of the consistency statistic for the example matrix are contained in Table 2.

In the example the consistency statistic is 612.85. For magnitude of consistency, the phi coefficient (e.g., Friedman, 1968) may be used. Here, the phi coefficient is the square root of the result obtained by dividing the consistency statistic by the total number of possible pairs. The phi coefficient has a maximum value of 1.00, which is obtained when the two sorts are identical, and a minimum value of .00, which is

Table 2
Observed and Expected Values for Determination of Chi-Square Statistic for the Example Matrix

Pair Type	Values	
	Observed	Expected
Trials 4 and 5	372.00	166.32
Trial 4 Only	135.00	340.68
Trial 5 Only	63.00	268.68
Neither Trial	756.00	550.32

obtained when there is no deviation of observed from expected values. The phi coefficient in the example is .68.

METHOD

Design

A 2 by 2 factorial design, with both factors varying between groups of subjects, was used. The first factor was the number of categories, either three or six, the subject was assigned to use in sorting. The second factor was the number of sorting trials, either 5 or 10, the subject was given before being asked to recall. The four resulting groups were labeled 3C-5T, 6C-5T, 3C-10T, and 6C-10T, with the first integer indicating the number of categories and the second integer indicating the number of trials.

Materials

The stimuli were 52 nouns randomly chosen from the Paivio, Yuille, and Madigan (1968) norms. Five randomizations of the nouns were tape-recorded for presentation. The intervals between words of a given randomization were 4 sec, and the intervals between randomizations were 10 sec. The same list and randomizations had been used in a previous study in which a significant correlation between number of categories and amount recalled was found when subjects were allowed to choose their numbers of categories (Schwartz & Humphreys, 1972, Experiment 3).

Answer booklets contained either 5 or 10 sorting pages, with lines for either three or six columns per page. The number of pages and the number of columns per page corresponded to the number of trials and the number of categories, respectively. The final page of each answer booklet contained 52 lines for recall.

Subjects

The subjects were 64 volunteers from introductory psychology classes at the University of British Columbia. The subjects were tested in four group sessions, and all of the subjects in a session received the same number of trials. The assignment of subjects to the number of categories variable was random within each session. There were 16 subjects in each of the four groups.

Procedure

The subjects were told that they would be presented 52 nouns, one at a time, and that their task was to put these nouns into categories. They were given answer booklets and were told to let each column represent a category. The number of categories they were to use, either three or six, was equivalent to the number of columns on their answer sheets. The subjects were told that they should sort according to the content of the nouns and that on each trial they should try to use the same organization as on the previous trial. They were informed of the number of sorting trials they would have and that they would be asked to recall the words at the end of the last sorting trial. They were allowed 3 min for recall. For subjects in the 10-trial groups, the five randomizations were repeated for the second five sorting trials.

RESULTS

The data of four subjects were eliminated from the analyses because the subjects did not follow instructions. Three of these subjects, all in Group 3C-10T, did not use content sorts; the fourth subject, in Group 6C-10T, did not use the number of categories to which he had been assigned. In addition, an a priori decision had been made to eliminate the data of subjects who were not attempting to sort consistently. An arbitrary standard for this decision was to eliminate data of subjects for which the consistency statistic for Trials 4 and 5 was less than 6.63. This would correspond to a p value of .01 if the consistency statistic was distributed as chi-square with 1 df. The lowest value for the consistency statistic was 13.00, so no data were eliminated for subjects' failure to attempt to sort consistently. The revised numbers of subjects in the four groups are contained in Table 3.

Recall, clustering, and consistency scores were determined for each subject. The mean recall scores for the four groups are contained in Table 3. The clustering scores reflect the amount of conformity of the subjects' recall protocols to the category structure used on the last sorting trial, and they are expressed in terms of z scores (see Frankel & Cole, 1971). Mean z scores for the four groups are also contained in Table 3. The consistency scores, expressed in terms of the phi coefficients relating Trial 4 and Trial 5 category structures were determined for all subjects, and phi coefficients relating Trial 9 and Trial 10 category structures were determined for the subjects in Group 3C-10T and Group 6C-10T. The means of phi coefficients are contained in Table 3.

In addition to recall, clustering, and consistency scores, Table 3 includes means for the four groups of the numbers of items which did not appear in the subjects' sorting protocols on Trials 4 and 5. The means for Groups 3C-10T and 6C-10T of the numbers of items which did not appear in the subjects' sorting protocols on Trials 9 and 10 are also included in Table 3. In

Table 3
Summary Data

Measure	Condition			
	3C-5T	6C-5T	3C-10T	6C-10T
Number of Subjects	16	16	13	15
Mean Recall	24.44	27.94	28.08	29.00
Mean Clustering	2.33	3.77	3.02	4.29
Mean Trial 4-5 Consistency	.76	.67	.81	.56
Mean Trial 9-10 Consistency			.98	.79
Mean Items Not Sorted*	1.06	3.44	1.15	4.53
Mean Items Not Sorted†			.23	2.20

**Trials 4 and 5* *†Trials 9 and 10*

general, the numbers of items not sorted were small; the largest mean shown in Table 3 of items not sorted represents only 4.4% of the total number of items presented. However, the six means of items not sorted were highly correlated with the six means of the consistency scores ($r = .87$, $p < .05$). Because of this high correlation, no further analyses of the numbers of items not sorted were made.

Two analyses were used to examine consistency scores. The first analysis examined consistency scores on Trials 4 and 5. The independent variables were the number of Categories, either three or six, and the number of Trials, either 5 or 10. The results indicated that subjects assigned to use three categories had higher consistency scores than did subjects assigned to use six categories [$F(1,56) = 9.54$, $MSe = .04$, $p < .01$]. The difference in consistency scores on Trials 4 and 5 between subjects given 5 sorting trials and those given 10 sorting trials was not significant ($F < 1$). Also, the interaction of Number of Categories by Number of Sorting Trials in Trials 4 and 5 consistency scores was not significant [$F(1,56) = 2.34$, $p > .10$].

The second analysis of consistency scores involved the data from subjects in Groups 3C-10T and 6C-10T. The independent variables were the number of Categories and the Trials, either Trials 4 and 5 or Trials 9 and 10, from which the consistency scores were measured. As expected from the results of the previous analysis, the subjects in Group 3C-10T had significantly higher consistency scores than did the subjects in Group 6C-10T, [$F(1,26) = 11.17$, $MSe = .06$, $p < .01$]. The improvement in consistency from Trials 4 and 5 to Trials 9 and 10 was significant [$F(1,26) = 40.29$, $MSe = .01$, $p < .01$]. The interaction of Number of Categories by Trials from which the consistency scores were measured was not significant ($F < 1$).

Analysis of variance was used to examine the recall scores. The independent variables were the number of Categories, either three or six, and the number of Trials, either 5 or 10. The results of the analysis indicated that subjects assigned to use three categories did not recall significantly fewer items than did subjects assigned to use six categories [$F(1,56) = 2.61$, $MSe = 32.53$, $p > .10$]. Also, subjects given 5 sorting trials did not recall significantly fewer items than did subjects given 10 sorting trials [$F(1,56) = 2.46$, $p > .10$]. The interaction of Number of Sorting Trials by Number of Categories was not significant ($F < 1$).

The clustering scores presented in Table 3 indicate that subjects in all groups were recalling according to their category structures; in fact, only one of the 60 recall protocols showed less clustering than would be expected by chance. Statistical analysis indicated that all groups were clustering at least at the .001 level, $z = 10.08$, 15.08, 10.89, and 16.62 for Groups 3C-5T, 6C-5T, 3C-10T, and 6C-10T, respectively.

Analysis of variance was used for further examination of the clustering scores. Number of Categories and number of Sorting Trials were the independent variables. The results of the analysis indicated that subjects assigned to use six categories had significantly greater clustering scores than did subjects assigned to use three categories [$F(1,56) = 14.64$, $MSe = 3.19$, $p < .01$]. Subjects given 10 sorting trials did not have significantly greater clustering scores than did subjects given 5 sorting trials [$F(1,56) = 1.97$, $p > .10$]. The interaction of Number of Categories by Number of Sorting Trials was not significant ($F < 1$).

All possible combinations of correlations between amount recalled, clustering scores, and Trials 4 and 5 consistency scores were determined for each of the four groups. In addition, correlations between amount recalled, clustering scores, and Trials 9 and 10 consistency scores were determined for Groups 3C-10T and 6C-10T. Table 4 contains the 18 resulting Pearson product-moment correlation coefficients.

The correlations shown in Table 4 present a fairly systematic pattern. Recall and clustering scores were significantly correlated only for Group 6C-10T, and recall and consistency scores were significantly correlated only for Group 6C-10T. Although none of the correlations between consistency and clustering was significant at the accepted level, they were all positive, and the correlation approached significance for Group 6C-10T.

The correlations between consistency scores on Trials 4 and 5 and on Trials 9 and 10 were positive for both Group 3C-10T and Group 6C-10T, although the correlation was significant only for Group 6C-10T. The failure to find a significant correlation between the two consistency scores for Group 3C-10T is probably due to the lack of variance in Trials 9 and 10

Table 4
Correlations Between Recall, Clustering, and Consistency

Correlation	3C-5T	6C-5T	3C-10T	6C-10T
Recall/Clustering	.10	−.30	.09	.69**
Recall/Trial 4-5 Consistency	.12	−.03	−.30	.67**
Recall/Trial 9-10 Consistency			.30	.81**
Clustering/Trial 4-5 Consistency	.22	.31	.02	.51*
Clustering/Trial 9-10 Consistency			.10	.49*
Trial 4-5 Consistency/Trial 9-10 Consistency			.38	.78**

*$p < .10$ **$p < .01$

consistency scores for that group; 9 of 13 subjects in Group 3C-10T had identical sorts on Trials 9 and 10. This is probably also the explanation for the nonsignificant correlation between Trials 9 and 10 consistency scores and recall for that group.

DISCUSSION AND CONCLUSIONS

Subjects who were instructed to sort into six categories were not as consistent in their sorting as were subjects who were instructed to sort into three categories. The former task is apparently more difficult, and it should take subjects longer to reach the same level of consistency as compared to those in the latter task. Thus, the correlations between number of trials and recall should be positive when calculated across the levels of the number of categories variable. There is also some indirect evidence which suggests that the correlation between number of trials and amount recalled might be negative when calculated within each level of the number of categories variable. The correlation between recall and Trials 9 and 10 consistency scores was significant for Group 6C-10T. The same correlation might have been significant for Group 3C-10T, except for the restriction on the range. Thus, the more consistent subjects, who should reach criterion faster on the sorting task, appear to be better recallers. It does not appear to be the case that they are better recallers because they are using their consistent organization to recall. If this were the case, the correlation between recall and clustering (there was no restriction on the range for the clustering measure) should have been higher for Group 3C-10T.

Apparently, the level of sorting consistency is not the explanation for Schwartz and Humphreys' (1972) failure to find a significant relationship between number of categories and amount recalled when both time per trial and number of trials were experimentally controlled. In the present experiment there was a significant increase in consistency from Trials 4 and 5 to Trials 9 and 10 (both the within-subject comparison and the between-subject comparison showed this increase). However, the magnitude of the category-recall function (the difference in the amount recalled between those subjects assigned to use six categories and those assigned to use three categories) was somewhat smaller after 10 sorting trials than it was after 5 sorting trials.

The failure to find a significant difference in recall between subjects given 10 sorting trials and those given 5 sorting trials was surprising. This failure may have resulted from a bias in assigning subjects to 5- and 10-trial groups (all of the subjects within a session were given the same number of trials). However, this possibility seems unlikely because subjects in 5- and 10-trial groups were comparable in terms of Trials 4 and 5 consistency scores. Instead, this failure to obtain a significant difference may reflect the fact that the learning curve is negatively accelerated and that it approaches an asymptote that is substantially less than 100% correct (see Humphreys & Schwartz, 1972, for a discussion of this issue with respect to the free recall paradigm).

The sorting task used in this experiment was designed by Mandler (1967) to accomplish specific goals. The task was supposed to allow subjects to impose or discover their own organization for a set of materials. Once this organization was imposed or discovered, it was supposed to be used by the subjects as a basis for recall. The results of the correlational analysis show that these objectives were obtained only for Group 6C-10T. The subjects in all four groups were organizing the words, as all of the subjects were consistent in their Trials 4 and 5 sorts. There was also a significant level of clustering in all four groups, but only in Group 6C-10T was there a significant correlation between the extent to which an individual subject clustered and the amount he recalled. Thus, it was only in Group 6C-10T that organization made a difference in the sense that those subjects who were best able to use their organization were the best recallers. The failure to find significant recall-clustering correlations in the five-trial groups may indicate that the process by which organization gets translated into recall takes time to develop. However, for 10-trial groups the only explanation appears to be that the subjects who were assigned to use six categories were learning or recalling in a different way than were those subjects who were assigned to use three categories.

When subjects have been randomly assigned to sort into either three or six categories and when time per trial and number of trials has been experimentally controlled, a small but positive category-recall function has been consistently found. In this study, after five sorting trials, subjects who were assigned to use six categories recalled on the average 3.5 more words than did subjects who were assigned to use three categories. After 10 trials this difference was .9 words. Schwartz and Humphreys (1972) found a difference of 1.7 words (Experiment 2) and 1.3 words (Experiment 3). This small but consistent difference in recall, as a function of whether the subjects were assigned to sort into six or three categories, need not reflect the organization per se. That is, the subjects who are assigned to the six-category task are presumably required to make finer discriminations than are the subjects who are assigned to the three-category task. The extra effort required, or perhaps the level of processing required (see Craik & Lockhart, 1972), may account for the superior performance in the six-category group.

In conclusion the category-recall function appears to have three components: (1) Subjects who choose to sort into six categories are better recallers than are subjects who sort into three categories. (2) Sorting into six categories is a more difficult task than sorting

into three categories, so subjects take longer to do the former task and presumably learn more. (3) The small effect left when subjects are randomly assigned to use three or six categories may be due to the organization or it may be due to the effort expended and/or the finer discrimination required.

REFERENCES

CRAIK, F. I. M., & LOCKHART, R. S. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 1972, **11**, 671-684.

FRANKEL, F., & COLE, M. Measures of category clustering in free recall. *Psychological Bulletin*, 1971, **76**, 39-44.

FRIEDMAN, H. Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin*, 1968, **70**, 245-251.

HUMPHREYS, M. S., & SCHWARTZ, R. M. The statistical evidence for negative transfer in part-whole free recall. *Behavior Research Methods & Instrumentation*, 1972, **4**, 287-291.

MANDLER, G. Organization and memory. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. I). New York: Academic Press, 1967.

MELKMAN, R. Effects of preferred and imposed number of categories on recall. *Journal of Experimental Psychology: Human Learning and Memory*, 1975, **1**, 280-285.

PAIVIO, A., YUILLE, J. C., & MADIGAN, S. A. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, 1968, **76**(1, part 2).

SCHWARTZ, R. M., & HUMPHREYS, M. S. Examinations of the category-recall function. *American Journal of Psychology*, 1972, **85**, 189-200.

(Received for publication January 5, 1976;
revision accepted April 1, 1976.)