Psychological Bulletin
1977, Vol. 84, No. 3, 539-556

The Measurement of Subjective Organization in Free Recall

Robert J. Sternberg
Yale University

Endel Tulving
University of Toronto

Alternative measures of subjective organization are presented and discussed. Various criteria for choosing among measures are compared, and four psychometric criteria are proposed: quantification, reliability, construct validity, and empirical validity. It is demonstrated that with respect to these criteria, the bidirectional form of intertrial repetition, here referred to as *pair frequency*, is the best measure of subjective organization available at the present time.

Multitrial free recall is a widely used task whose analysis has been thought to provide important insights into processes of learning and memory. In a typical multitrial free-recall task, the learner is shown a list of single words in the study (input) phase of a trial, and he is asked to recall as many of these words as he can remember in the recall (output) phase of the trial. The learner is permitted to recall the words in any order he wishes; hence the designation *free recall*. The same set of to-be-learned words is shown on a number of successive trials, always in a different order; hence *multitrial free recall*.

At least two things happen in the course of multitrial free recall of a list of words: (a) The number of words recalled increases over trials, that is, the learner learns the list, and (b) the order of words recalled becomes increasingly stereotyped, that is, the learner organizes his recall. Since the order of words presented for study varies unsystematically from trial to trial, the increasing sequential organization of words over trials must be imposed upon the material by the learner; hence its label *subjective organization* (Tulving, 1962).

The two basic empirical phenomena of multitrial free recall, learning and subjective organization, have constituted the source of a number of questions to which answers have been sought in research. Why does the number of words recalled increase over trials? Why do learners organize their recall, even though it is not a part of the task requirements? What is the nature of the relation between free-recall learning and subjective organization? Are they both parallel manifestations of the same set of underlying processes, or does one somehow cause the other? These and other similar questions have occupied the minds of students of verbal learning and memory for some time now and, as no consensus on the answers has emerged, they are likely to be with us in the future.

If we wish to understand free-recall learning and subjective organization, and especially the relation between them, we must be able to discuss them in quantitative terms. The problem of measurement seems to have been satisfactorily resolved for learning, inasmuch as most experimenters and theorists agree that the simple number or proportion of words recalled represents a suitable measure of recall performance. Learning is then defined in terms of change in recall over trials. The problem of measurement is as yet unsolved, however, for subjective organization. A number of different measures of subjective organization have been proposed and are currently being used, but very little is known about their relations, about the reasons why any particular measure is used in

This research has been supported by grants from the National Research Council of Canada and the National Science Foundation to Endel Tulving. It was partly written during Endel Tulving's residence as a fellow at the Center for Advanced Study in the Behavioral Sciences, Stanford, California.

The authors are grateful to Krystyna Dronsejko for experimental assistance, to Carol Treanor for computational help, and to Phipps Arabie, Gerald Balzano, Michael Friedman, and George Mandler for many helpful comments on the manuscript. The authors especially thank Mary W. Laurence for making the raw data from her experiment available to us.

Requests for reprints should be sent to Robert J. Sternberg, Department of Psychology, Box 11A, Yale Station, Yale University, New Haven, Connecticut 06520.

a particular experiment, and about whether any one or more of the existing measures are somehow better than the others, and if so, why.

In the absence of a comparative analysis of measures of subjective organization, the choice of measures by an experimenter is necessarily arbitrary, interexperimental comparisons of findings are difficult, and progress in understanding phenomena of multitrial free recall may be retarded. It is quite conceivable that different experimenters can reach different conclusions about subjective organization in multitrial free recall, simply because they use different ways of quantifying organization. If these different ways of quantifying organization are equally valid, reliable, and meaningful, then the resolution of the conflict must be sought on other grounds. If, on the other hand, some ways of measuring organization are more appropriate than others, then the conflict may be more apparent than real; it will vanish when the measurement problem has been solved.

The purpose of the present review is to compare and evaluate systematically the measures of subjective organization that exist in the literature. We describe extant measures, briefly discuss their rationale, and illustrate their nature and differences with the aid of a concrete, albeit hypothetical, set of free-recall protocols. We then discuss the criteria that seem appropriate for the evaluation of measures of subjective organization, and compare six of the most popular measures in light of data from an extensive multitrial free-recall experiment. We conclude that one of the measures, a variant of the intertrial repetition measure, first proposed by W. A Bousfield (A. K. Bousfield & W. A. Bousfield, 1966; W. A. Bousfield, Puff, & Cowan, 1964), has certain advantages over the others and that it should therefore be regarded as the measure of choice. Finally, we illustrate, by means of an actual example, how the use of the preferred measure of subjective organization can drastically change the conclusions about the relation between learning and organization arrived at by the use of a less adequate measure.

Measures of Subjective Organization

Theoretical Basis

Subjective organization, like many other terms in psychology, refers to two different, albeit closely related, concepts. One is a psychological process; the other is a measure of the extent to which the process is revealed in observable behavior.

The process notion of subjective organization derives from Miller's (1956a, 1956b) concept of chunking, or unitization. It is assumed that a subject's free-recall capacity is limited to a relatively small number of chunks, or subjective units, of material (S units). This limit in free recall is (a) independent of the size of the units and (b) relatively constant over successive trials (Tulving, 1964). When the subject studies the list, he groups (organizes) more and more individual list items into higher order S units; when he recalls the list, he retrieves S units one at a time and produces the constituent words of each in succession.

Two observable consequences follow directly from this organizing activity: The number of words recalled increases over trials (even though the number of S units that can be recalled remains unchanged), and words from the same higher order unit are recalled either in adjacent output positions or at least in close temporal proximity to one another. One might argue that the former phenomenon is explained by the latter; according to the theory, however, both phenomena are manifestations of one and the same underlying process. To measure subjective organization usually means to measure the extent to which the output order of words is sequentially constrained over successive trials, since the increase in recall over trials could be accounted for by processes other than subjective organization. But constancies in output order, under conditions wherein such constancies are not required of subjects' performance, cannot be readily attributed to processes other than unitization of elementary units into higher order ones. The degree of output consistency over trials can thus be used as an index of the extent to which a particular organization has occurred and is maintained from one trial to the next (Tulving, 1962).

An Imaginary Free-Recall Protocol

Consider for example a situation in which a person is given a list of 12 words to learn over 6 trials. The list consists of the following words: *palace, bank, poem, demon, sorrow, forest, game, walnut, unicorn, king, joke, mountain.* Words are presented via the typical multitrial free-recall procedure. A possible outcome of the experiment for a single subject is depicted in Table 1, which lists the words the subject recalled on the 6 trials, in the order in which the subject recalled them. Table 1 also shows certain numerical data relevant to measures of subjective organization; we will have occasion to refer to these data as we go along.

Table 1
Illustrative Data from an Imaginary Multitrial Free-Recall Experiment: Words Recalled by a Subject on Six Trials

Output position	1	2	3	4	5	6
1	king	walnut	forest	demon	joke	poem
2	palace	sorrow	game	bank	forest	mountain
3	forest	mountain	bank	mountain	poem	forest
4	poem	game	mountain	poem	mountain	unicorn
5		forest	walnut	forest	demon	king
6		poem	sorrow	game	bank	demon
7				unicorn	walnut	bank
8					sorrow	walnut
No. correct (R)	4	6	6	7	8	8
Overlapping words (c)	—	2	5	4	5	6
Forward pairs	—	1	1	2	1	3
Backward pairs	—	0	1	0	2	0
Forward triplets	—	0	0	0	0	1
Backward triplets	—	0	0	0	1	0
Unordered triplets	—	0	1	0	1	1

The recall protocol in Table 1 manifests both an increase in recall over trials (learning) and a certain amount of consistency in the order in which the words are recalled on different trials (subjective organization). The number of words recalled (R) increases from 4 to 8 in the course of the 6 trials, and even a casual inspection suggests that the ordering of words tends to be quite similar from one trial to the next, particularly on later trials. Measures of subjective organization differ in the way in which they quantify this similarity. We next turn to a consideration of these measures as they developed historically.

The Subjective Organization (SO) Measure

Unidirectional SO. The first output adjacency measure of subjective organization to be proposed was the SO measure (Tulving, 1962). The measure is an information-theoretic one and is computed on the basis of a matrix in which all presented words are placed along both rows and columns. The rows represent Word i recalled in a given output position, and the columns represent Word j recalled in the next position. An additional "entry" row (row 0) and "exit" column (column 0) are also used, corresponding to an imaginary item at the beginning and at the end of the recall protocol. Cell entries (n_{ij}) in the matrix are the frequencies with which recall of Word j followed recall of Word i in the block of trials under consideration. Marginal entries (n_i and n_j) are summed frequencies of rows and columns. Table 2 shows the matrix for the imaginary protocol presented in Table 1.

The SO measure is defined[1] as

$$SO = \frac{\sum_{ij} n_{ij} \log n_{ij}}{\sum_{i} n_i \log n_i}. \quad (1)$$

In Equation 1, n_{ij} represents the numerical value of the cell in the ith row and jth column, and n_i represents the marginal total for the ith row. The value of SO for the imaginary protocol in Table 1 is .30 for the block of 6 trials.

Bidirectional SO (SO2). A bidirectional form of SO has also been used in computing subjec-

[1] This definition is for the usual case in which the number of trials does not exceed the number of words in the list. See Tulving (1962) for the general case.

Table 2
Illustrative Recall Matrix for Computation of SO: Data Over Six Trials for Subject in Imaginary Experiment

	(n + 1)th word													
nth word	0	1	2	3	4	5	6	7	8	9	10	11	12	n_i
0				1		1		1			1	1	1	6
1						1								1
2					2							2		4
3	2				2	1								5
4				1		1		1	1			1		5
5				3					2	1				6
6	2				1									3
7			3											3
8			1			1				1				3
9	1												1	2
10						1								1
11	1						3							4
12		1						1						2
n_j	6	1	4	5	5	6	3	3	3	2	1	4	2	45

Note. Word equivalents: 1 = palace, 2 = bank, 3 = poem, 4 = mountain, 5 = forest, 6 = sorrow, 7 = demon, 8 = game, 9 = unicorn, 10 = joke, 11 = walnut, 12 = king.

tive organization (e.g., Gorfein & Blair, 1971). In calculating this measure, SO2, one takes into account both forward and backward repetitions. In the imaginary protocol of Table 1, for example, the word pair *game, forest* appears in Trial 2, and the word pair *forest, game* appears in Trial 3. This pair would be counted as a repetition in SO2, but not in standard (unidirectional) SO. The formula for SO2 for the simplified case in which the number of trials does not exceed the number of list words is

$$SO2 = \frac{\sum_{ij} (n_{ij} + n_{ji}) \log (n_{ij} + n_{ji})}{2 \sum_{i} n_i \log (n_i)}. \quad (2)$$

The value of SO2 for the imaginary protocol in Table 1 is .23.

The Intertrial Repetition (ITR) Measure

Unidirectional ITR. An obvious problem with SO (and SO2) is that there is no baseline for chance organization. Even with randomly generated protocols, some organization would appear, due to chance adjacencies in the recall data. Since the expected value of SO is not known, there is no mathematical correction for chance. An output adjacency measure of subjective organization has been proposed by W. A. Bousfield (A. K. Bousfield & W. A. Bousfield, 1966; W. A. Bousfield, Puff, & Cowan, 1964) that does have a known expected value. This measure, ITR, is defined as

$$ITR = O(ITR) - E(ITR), \quad (3)$$

where O(ITR) is the observed and E(ITR) is the expected value of the number of intertrial repetitions. An intertrial repetition is a pair of items recalled in two adjacent output positions on Trial t and recalled in two adjacent output positions and in the same order on Trial $t + 1$. The observed value of ITR is equal to the number of such intertrial repetitions that occur on a given pair of trials. For example, the successive pair of words, *forest, game*, in Output Positions 1 and 2 on Trial 3 also appears on Trial 4 in Output Positions 5 and 6. This pair thus adds a score of 1 to the value of O(ITR) on Trials 3 and 4. Words *bank, mountain* add another score of 1, for a total value of 2 for O(ITR) on these two trials.

The expected value of ITR, according to A. K. Bousfield and W. A. Bousfield (1966), is

$$E(ITR) = \frac{c(c-1)}{hk}, \quad (4)$$

where c is the number of common items recalled on both Trials t and $t + 1$, h is the number of items recalled on Trial t, and k is the number of items recalled on Trial $t + 1$. For Trials 3 and 4, for example, the value of E(ITR) is $4(4 - 1)/(6 \times 7) = 12/42 = .286$. This formula for chance expectation is based upon the "assumption that the given recall sequence is a random sample from among all orderings of the recalled items" (A. K. Bousfield & W. A. Bousfield, p. 939).

The ITR measure, unlike SO, is always computed for a block of just two trials. Its value (after correction for expected value) may range from $-E(ITR)$ to $c - 1 - E(ITR)$. The values of ITR for the imaginary protocol are .92, .44, 1.71, .64, and 2.53 on the five successive blocks of two trials (Trials 1 and 2, 2 and 3, 3 and 4, 4 and 5, 5 and 6).

Bidirectional ITR (pair frequency). ITR, like SO, can be computed in bidirectional form. The bidirectional form of the measure has been used by Anderson and Watts (1969) and by Rosner (1970). We shall call this measure *pair frequency* (PF), in order to prevent confusion with a number of similar sounding ITR-like measures to be described shortly. PF, like ITR, is a difference measure and is represented by

$$PF = O(ITR2) - E(ITR2) = O(ITR2) - \frac{2c(c-1)}{hk}. \quad (5)$$

In Equation 5, O(ITR2) represents the number of pairs of items recalled on Trials t and $t + 1$ in adjacent output positions in either of two possible orders, E(ITR2) represents the expected number of pairs of items, and c, h, and k retain the same meaning as in Equation 4. For example, one forward pair is common to Trials 2 and 3 (*walnut, sorrow*), as well as one backward pair (*forest, game*), for a total O(ITR2) score of 2. The E(ITR2) score is 1.11. Hence, PF is .89 for this block of two trials. The values of PF for the imaginary protocol are .83, .89, 1.43, 2.29, and 2.06 across the five successive pairs of trials.

Generalized ITR. An apparent shortcoming of all the measures described so far is that they measure subjective organization only for pairs of words. Mandler (1967) and Postman (1972), among others, have criticized the measures for inability to reflect higher order units of organization. Pellegrino (1971) has generalized the ITR measure to handle units of arbitrarily large size. Consider, for example, triplets. In the imaginary protocol, a forward triplet is repeated between Trials 5 and 6 (*demon, bank, walnut*). One may therefore count it as a higher order ITR. As with simple ITR, one may count backward repetitions of triplets in addition to forward ones. In the imaginary protocol, the triplet *mountain, poem, forest* that appears in Trial 4 appears in backward order in Trial 5 (*forest, poem, mountain*). Units may also be counted without regard to order, allowing for $n!$ possible permutations of any n-tuplet. For example, the triplet *forest, poem, mountain* appears in Trial 5, whereas *poem, mountain, forest* appears in Trial 6. This set of three items would be counted as an unordered unit of Size 3.

Pellegrino (1971) has also presented a formula for the expected value of ITR that is general to all word-unit sizes and directions:

$$E(ITR) = \frac{(N - X + 1)!(A)(M - X + 1 - R)}{N!}, \quad (6)$$

where M is the number of items recalled on Trial t, N is the number of items recalled on Trial $t + 1$, X is the size of the subjective organization unit, R is the number of units of Size X from Trial t that have one or more items not recalled on Trial $t + 1$, and A is the variable parameter dependent upon the specific order within the subjective organization unit ($A = 1$ for unidirectional units, $A = 2$ for bidirectional units, and $A = X!$ for unordered units).

This formula differs slightly in its assumptions from A. K. Bousfield and W. A. Bousfield's (1966) formula (see Pellegrino, 1971), and so it gives different results. For example, the values of unidirectional ITR for word pairs are .83, .33, 1.57, .50, and 2.38, which differ from the values presented previously, which were computed on the basis of Bousfield and Bousfield's formula for expectation. The comparable values for unidirectional triplets are lower, reflecting the reduced occurrence of larger units in the imaginary and in real protocols: .00, −.10, −.05, −.05, .93.

Generalized adjusted ratio of clustering (ARC'). Pellegrino (1971) criticized difference

now increases monotonically as a function of age.

Rosner's discounting of the difference between SO and ITR as a possible source of the discrepancy between her data and Laurence's is understandable in light of her reliance upon the correlations between measures reported by Puff and Hyson (1967). These authors reported correlations between SO and ITR of .94 and .97. Their subjects (college and nursing students) learned a list of 10 words over 20 trials. Although the authors did not report mean levels of recall on this task, it is reasonable to assume that most subjects attained perfect or near-perfect performance after only a few trials. It is known that the correlation between organization and recall increases over trials (Shapiro & Bell, 1970; Tulving, 1964), and it is possible that the extremely high correlations obtained by Puff and Hyson can be understood in these terms. An examination of their Figure 1 will reveal that in the early trials of their experiment, probably before recall reached asymptote, the correlation between mean SO and ITR was not nearly as high as it became in later trials.

Our reanalysis of the Laurence data suggests that the results of this study were puzzling only because of the use of SO as a measure of subjective organization. The pattern of results obtained with PF is theoretically more meaningful and is consistent with the data obtained independently by Rosner (1971). One can only speculate as to how many other experimental outcomes have been rendered less interpretable, or even subject to misinterpretation, because of the use of an inferior measure of subjective organization.

Conclusions

The evidence accumulated to date strongly suggests that under our criteria and theory of subjective organization described above (Tulving, 1962), the best available measure of subjective organization is pair frequency, or PF (Equation 5). Even if one accepts only the classical-test-theory notion of true score and its relationship to reliability (and ignores the empirical validity data), PF still comes out as the preferred measure. To question our conclusion, one must question our criteria, from which our conclusion follows directly.

Is there really such a thing as a "best" measure of subjective organization? In discussing a related type of measurement, the measurement of categorical clustering, Shuell (1975) concluded that "the best strategy to follow is to realize that there is no such thing as the best measure" (p. 723). His answer to the question, "Which measure of clustering or organization is the best one to use?" is another question, "Best for what?" (p. 723).

A more appropriate question, in our view, is "Best in what sense?" It is in the answer to this question that we have found the most disagreement and confusion to have arisen. Investigators have known what they wanted to measure, but not in what sense some measures are better than others. We have attempted to provide a clear answer to this question: best in the sense that the adopted measure most adequately meets certain explicitly stated criteria. We have proposed four such criteria: quantification, reliability, construct validity, and (nonartifactual) empirical validity. One measure, PF, has shown itself superior on these criteria, and we therefore believe that of the measures we have tested, it is indeed the "best."

Our conclusion is not intended to be the final word in the measurement of subjective organization. Important theoretical questions regarding the nature of the organizing process remain unanswered, and it is quite possible that organization theory and measurement will continue to interact, so that advances in one pursuit will lead to developments in the other. Eventually, a theory of subjective organization may be so well specified that it permits only one measure. Then, of course, there will be no problem of selecting from among alternative measures. At the present time, no such theory exists. For the present, if one's purpose is the modest one of measuring subjective organization rather than of reconceptualizing it in terms of some new theory, then one cannot do better than to measure it with PF.

Reference Note

1. Friendly, M. L. *Proximity analysis and the structure of organization in free recall* (ETS Research Bulletin RB-72-3). Princeton, N. J.: Educational Testing Service, 1972.

References

Anderson, R. C., & Watts, G. H. Bidirectional associations in multi-trial free recall. *Psychonomic Science*, 1969, *15*, 288–289.