Behavior Research Methods & Instrumentation
1974, Vol. 6 (4), 385-388

METHODS & DESIGNS

Comparison of output order in free recall*

LUIS M. FLORES, JR.
Kansas State University, Manhattan, Kansas 66502

and

SAM C. BROWN†
University of Missouri, Columbia, Missouri 65201

Two measures of output order in free recall were examined and shown to vary with characteristics of recall unrelated to output order. The output location of a subset of items in recall, using the standard recall rank measure and the "observed minus expected" score, varied both with number of items in the subset and total number of items recalled. A new measure of output order (the relative index of priority or RIP score) was proposed that is invariant with these characteristics of recall, providing an uncontaminated empirical index of the output location of a subset of items in the recall sequence.

Numerous investigations of human memory have been concerned with retrieval schemes that Ss employ during the recall of verbal material. Much of this research has involved the free-recall paradigm, because this procedure gives Ss the opportunity to recall items in any order they choose. Most notable of the retrieval schemes studied thus far have been the tendency for taxonomically related items to be recalled adjacent to each other (i.e., clustering; see Shuell, 1969) and the tendency for "unrelated" words to be recalled together on successive trials (i.e., subjective organization; see Tulving, 1962). Both clustering and subjective organization, however, are concerned only with the *grouping* of items together in recall. Another aspect of recall which has received less attention, but which also has provided important information about retrieval strategies, is the *order* of recall of items. In 1965, Battig, Allen, and Jensen reported that Ss tend to emit first in the recall sequence items recalled correctly for the first time (new items); they then emit items that had been recalled on one or more previous trials (old items). The tendency for Ss to recall new before old items has been labeled the priority effect. The priority effect has been replicated and extended and has led to the formulation of several hypotheses concerning output order in free recall (e.g., Battig & Slaybaugh, 1969; Brown & Thompson, 1971; Mandler & Griffith, 1969; Roberts, 1969).

Because the theoretical generalizations derived from the examination of recall protocols rest heavily on the validity of the quantitative measures applied to empirical data, the question of measurement has been of major concern. Recently, Roenker, Thompson, and Brown (1971) discussed the difficulties of previous measures of clustering and developed a measure that was free of the limitations of its predecessors. Likewise, Pellegrino (1971) reviewed existing measures of intertrial repetitions and derived a method for the measurement of higher-order subjective organization. The purpose of the present paper is to examine measures of output order and to propose a new measure that is free of some of the limitations of existing measures.

Previous investigations have employed essentially one of two measures of output order. One is the standard recall rank (SRR) score developed by Battig et al (1965), and the other is an "observed minus expected" (O-E) difference score introduced by Postman and Keppel (1968) and Shuell and Keppel (1968). These measures have been used exclusively to measure priority. Accordingly, we will discuss and compare measures of output order in the context of priority, although the measure we will propose can be used to determine the location of any item in the recall protocol (e.g., items presented in end or other input positions, items that differ in rated meaningfulness, and so forth). Before proceeding further, it is important first to specify how the SRR and O-E scores are defined.

THE SRR SCORE

For a given recall protocol, an SRR score for each item is obtained by the following relationship:

$$\mathrm{SRR} = \frac{\mathrm{Mdn}_R - R_i}{\sigma_R}, \qquad (1)$$

where R_i = Item i's output rank position with the item

---

*Based on an MS thesis by the first author at Kansas State University. The authors thank Charles P. Thompson, William F. Battig, and James W. Pellegrino for their suggestions for improvement of the manuscript.

†Requests for reprints should be addressed to Sam C. Brown, Department of Psychology, University of Missouri, Columbia, Missouri 65201.

Table 1

Quarter of Recall	Output Rank	Recall Protocol A	B	C	D	E	SRR
	1	n	n	n	n	n	1.59
First	2	n	n	n	n	o	1.30
	3	n	n	n	o	o	1.01
	4	n	n	o	o	o	.72
Second	5	n	o	o	o	o	.43
	6	o	o	o	o	o	.14
Third and Fourth	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	12	o	o	o	o	o	−1.59
O-E		1.75	2.00	2.25	1.50	0.75	
$\overline{\text{SRR}}$		1.01	1.16	1.30	1.44	1.59	

recalled first assigned a rank of 1, the second, 2, etc.; Mdn_R = median output rank for all i items; and σ_R = standard deviation of the total number of ranks. The SRR score is essentially a z score, in that each item is expressed in terms of its distance from the median in standard deviation units. Items recalled above the median rank take on positive values and those below, negative values. The algebraic mean of the deviations for new items ($\overline{\text{SRR}}$) constitutes the index of priority.

THE O-E SCORE

The second measure of priority is the difference between the observed and expected number of new items occurring in some specified segment of the recall protocol, usually the first quarter. If new items occur randomly in different positions of the output sequence, then one-fourth of the total new items recalled would be expected to occur in each quarter of the protocol. If the O-E difference is positive, then more new items are recalled in the first quarter than expected by chance.

LIMITATIONS OF THE $\overline{\text{SRR}}$ AND O-E SCORES

Both measures suffer from the limitation that they vary with characteristics of recall unrelated to relative amounts of priority. Specifically, the score for maximum positive priority (all new items recalled before old items) and maximum negative priority (all new items recalled last) changes as a function of both the total number of new items recalled (N) and the total number of items recalled (T). Each will be discussed in turn.

Number of New Items Recalled

Consider the case when all new items are recalled first (i.e., maximum positive priority). The five hypothetical recall protocols (A-E) in Table 1 are all examples of maximum positive priority. All protocols are of the same length (12), but differ in number of new (n) and old (o) items recalled. The SRR score for each item is given in the last column. The scores are the same for items with identical output ranks because each protocol is of the same length. The size of a quarter of recall for all protocols is three (Q = 3), and the number of new items recalled ranges from one to five. As can be seen from the bottom two rows, both O-E and $\overline{\text{SRR}}$ scores for new items yield different values for protocols of varying numbers of new items recalled even when all of the new items are emitted before old items. In the case of O-E, the score for maximum positive priority will be highest when N = Q (Protocol C). Increasing N beyond Q (Protocols A and B) or decreasing N below Q (Protocols D and E) yields progressively smaller O-E scores. The same biases apply to maximum negative priority, but in the opposite direction because maximum positive and maximum negative priority are symmetric to each other. Also, the present objections to O-E apply when any segment of the recall protocol is used as the unit of analysis (e.g., fourths, halves, or eighths of recall).

The relationship between $\overline{\text{SRR}}$ and number of new items is different. Here, $\overline{\text{SRR}}$ increases as N decreases. This occurs because $\overline{\text{SRR}}$ is the average of the SRR scores for all new items. Decreasing N while maintaining maximum positive priority serves to eliminate the lower SRR scores (of individual items), and therefore results in a larger $\overline{\text{SRR}}$ score.

Total Number of Items Recalled

In the case of the O-E measure, holding N constant while varying T will yield the same O-E score only if $N \leqslant Q$ in each recall protocol. Consider Protocols F and G in Table 2. In each instance $N \leqslant Q$. Hence, both protocols produce the same O-E score. Now consider Protocols H and I in Table 3. In this situation $N > Q$ for Protocol H. The O-E score for Protocol H is 1.25, while for Protocol I it is 2.25. Thus, the O-E score for maximum positive priority will be the same for protocols of different lengths, if for all protocols $N \leqslant Q$. When $N > Q$, O-E scores will differ from one another in a manner related to the magnitude of the difference between N and Q: the greater the difference between N and Q, the lower the O-E score.

The $\overline{\text{SRR}}$ measure also fails to yield the same score for maximum positive priority when the lengths of protocols differ. Given two protocols with the same number of new items in maximum positive priority positions, $\overline{\text{SRR}}$ will be larger for the protocol with the larger total number of items recalled (T). This is illustrated by comparing Protocol F with G and

Table 2

	Quarter 1 2 3 4	Q	N	O-E	$\overline{\text{SRR}}$
Protocol F	nn/ oo/ oo/ oo	2	2	1.50	1.30
Protocol G	nno/ooo/ooo/ooo	3	2	1.50	1.44

Table 3

	Quarter 1 2 3 4	Q	N	O-E	$\overline{\text{SRR}}$
Protocol H	nn/ no/ oo/ oo	2	3	1.25	1.09
Protocol I	nnn/ooo/ooo/ooo	3	3	2.25	1.30

Protocol H with I. In both instances, the protocol with the greater number of total items recalled has the higher $\overline{\text{SRR}}$ score. The reason can be understood by referring to Formula 1. As list length increases, both the numerator ($\text{Mdn}_\text{R} - \text{R}_\text{i}$) and the denominator ($\sigma_\text{R}$) of Formula 1 will also increase (by a constant value). However, with the addition of each old item to recall, the numerator increases by a greater value (.50) than the denominator (.29). Therefore, $\overline{\text{SRR}}$ will increase as T increases, although the magnitude of the increase will become smaller as more old items are added to recall. Protocols F and I yield the same $\overline{\text{SRR}}$ score even though they differ in length and number of new items. This will occur for maximum priority (positive or negative) whenever the ratio of new to old items is the same in different protocols. Parenthetically, it is of interest to mention that previous investigations using the $\overline{\text{SRR}}$ measure have reported an increase in priority for new items over trials (e.g., Battig et al, 1965; Brown & Thompson, 1971). As pointed out above, $\overline{\text{SRR}}$ increases as T increases and as N decreases. Since both an increase in T and a decrease in N would be expected over the normal course of multitrial free recall learning, previous reports of increasing priority could be an artifact of the $\overline{\text{SRR}}$ measure.

THE RELATIVE INDEX OF PRIORITY (RIP) SCORE

To overcome the shortcomings of the $\overline{\text{SRR}}$ and O-E measures, we have developed a new priority measure, the relative index of priority (RIP) score. Basic to RIP is the specification of an item's location in the recall protocol in terms of its output rank position. Thus, the sum of the output ranks for new items (observed priority) should be related to the level of priority in a given protocol. If there is no priority, then new items should be equally scattered above and below the median output rank (i.e., the sum of the algebraic ranks of new items from the median output rank should equal zero). In RIP, the condition of no priority is indexed by the sum of the ranks of the N middle positions. Likewise, maximum positive priority is indexed by the sum of the first N ranks, and maximum negative priority by the sum of the last N ranks. For example, consider a protocol where the total number of items recalled is nine and the number of new items recalled is three. If the first item recalled is assigned a rank of 1, the second a rank of 2, and so on, then maximum positive priority, no priority, and maximum negative priority would equal 6, 15, and 24, respectively. What RIP does is to express the difference between observed priority and no priority as a proportion of the maximum possible difference from no priority. The definitional formula for RIP is given by:

$$\text{RIP} = (\Sigma \text{R}_\text{i} - \text{NP})/(\text{MaxP} - \text{NP}), \tag{2}$$

where $0 < \text{N} < \text{T}$, and ΣR_i = observed priority which is the sum of the rank output positions of new items; MaxP = maximum positive priority, i.e., sum of rank positions if all new items were recalled first; NP = no priority, i.e., sum of ranks if all new items were emitted in the middle positions of a protocol; N = total number of new items recalled; and T = total number of items recalled.

It is important to note that for maximum positive priority, no priority, and maximum negative priority, Formula 2 will always yield scores of 1, 0, and −1, respectively. This is true for any protocol regardless of its value for N and T. For example, Protocols A-E and F-I will yield in each instance a RIP score of 1. Had all new items in each protocol been recalled last, the RIP score would be −1.[1]

A computational formula for RIP is easily derived by the following relation, which is the sum of an arithmetic series:

$$\text{MaxP} = \text{N}(\text{N} + 1)/2. \tag{3}$$

Further, from Formula 3, it can be shown that

$$\text{NP} = \text{N}(\text{T} + 1)/2. \tag{4}$$

Substituting Formulas 3 and 4 in Formula 2 and simplying results in the following computational formula:

$$\text{RIP} = [\text{N}(\text{T} + 1) - 2\Sigma \text{R}_\text{i}]/\text{N}(\text{T} - \text{N}), \tag{5}$$

where $0 < \text{N} < \text{T}$, and ΣR_i, N, and T are defined as in Formula 2.

To summarize, because of its fixed upper and lower bound (regardless of the value for N and T), RIP has a distinct and powerful advantage over other measures of priority: it allows for comparison of relative amounts of priority across trials for a single S, as well as for comparisons between Ss and between experiments. The RIP score therefore has the same basic properties as the ARC score for measuring clustering (Roenker, Thompson, & Brown, 1971) and its extension for measuring "subjective" and input-output organization (Pellegrino, 1971). This makes it possible to compare directly the relative magnitudes of each of these measures of list organization in free recall learning.

OUTPUT POSITION OF A SINGLE ITEM

When the output location of an individual item is of concern, Formula 2 becomes:

$$\text{RIP}_\text{i} = (\text{T} + 1 - 2\text{R}_\text{i})/(\text{T} - 1), \tag{6}$$

where i = 1, 2, 3, ···, T, and R_i = Item i's output rank position. Formula 6 has all the properties of SRR with the added advantage of having fixed upper and lower bounds equal to 1 and −1, respectively, for protocols with any values of N and T (where $0 < \text{N} < \text{T}$). Thus,

Table 4

	Output Rank Position										
	1	2	3	4	5	6	7	8	9	10	RIP
Protocol J	o	n	o	n	o	o	n	o	n	o	0
Protocol K	n	n	o	o	o	o	o	o	n	n	0

when Item i has a RIP_i score of 1, this indicates that it was the first item emitted, while -1 indicates that it was the last item emitted, etc.

Individual priority scores may be useful when, for example, the output location of new items presented in recency positions is of concern. One could entertain the hypothesis that new items presented in recency positions during study will be output before other new items (e.g., Brown & Thompson, 1971). For each protocol, a mean RIP_i score could be calculated for new items presented in recency positions and compared with the mean RIP_i score for other new items. This procedure, however, may result in the same biases as discussed in relation to $\overline{SRR}$: RIP_i scores averaged in this manner do not express the actual amount of priority as a proportion of the total priority possible. To avoid these biases, RIP scores could be computed separately for new items presented in recency and nonrecency positions.

MEANING OF THE RIP SCORE

The problems involved in the interpretation of the RIP score generally apply to the interpretation of any statistic. Consider, for example, Protocols J and K in Table 4. Both protocols contain four new items, the ranks of which sum to 22. Both, therefore, yield a RIP score of zero. However, the distribution of new items in each protocol is different. In Protocol J, new items are distributed evenly in the recall sequence, while in Protocol K they are grouped at the beginning and end of the output sequence. Clearly, the RIP score does not differentiate between these two protocols. It therefore behooves the investigator to examine individual protocols for systematic patterns of recall such as in Protocol K. If the data indicate, for example, that Ss are emitting some new items at the beginning and others at the end of the output sequence, then separate RIP scores for the first and second half of recall may be computed. Other cases may of course warrant different applications of RIP.

CONCLUSION

The RIP score is free of the limitations of previous indices and provides an uncontaminated measure of relative amounts of priority in free recall. Moreover, the RIP measure may be applied to any type of item and is much easier to compute than other measures. It should be emphasized that RIP provides *only* an empirical index of the relative output location of a subset of items in recall. It does not specify the mechanisms or S strategies underlying the recall order. Nonetheless, we believe that such specification will rest ultimately upon the use of a measure which accurately describes the phenomenon to be explained.

REFERENCES

Battig, W. F. Allen, M., & Jensen, A. R. Priority of free recall of newly learned items. Journal of Verbal Learning & Verbal Behavior, 1965, 4, 175-179.

Battig, W. F., & Slaybaugh, G. D. Evidence that priority of free recall of newly learned items is not a recency artefact. Journal of Verbal Learning & Verbal Behavior, 1969, 8, 556-558.

Brown, S. C., & Thompson, C. P. Relationship between item strength and order of free recall. Journal of Verbal Learning & Verbal Behavior, 1971, 10, 444-448.

Mandler, G., & Griffith, M. Acquisition and organization of new items in the free recall of random lists. Journal of Verbal Learning & Verbal Behavior, 1969, 8, 545-551.

Pellegrino, J. W. A general measure of organization in free recall for variable unit size and internal sequential consistency. Behavior Research Methods & Instrumentation, 1971, 3, 241-246.

Postman, L., & Keppel, G. Conditions determining the priority of new items in free recall. Journal of Verbal Learning & Verbal Behavior, 1968, 7, 260-263.

Roberts, W. A. The priority of recall of new items in transfer from part-list learning to whole-list learning. Journal of Verbal Learning & Verbal Behavior, 1969, 8, 645-652.

Roenker, D. L., Thompson, C. P., & Brown, S. C. A comparison of measures for the estimation of clustering in free recall. Psychological Bulletin, 1971, 76, 45-48.

Shuell, T. J. Clustering and organization in free recall. Psychological Bulletin, 1969, 72, 353-374.

Shuell, T. J., & Keppel, G. Item priority in free recall. Journal of Verbal Learning & Verbal Behavior, 1968, 7, 969-971.

Tulving, E. Subjective organization in free recall of "unrelated" words. Psychological Review, 1962, 69, 344-354.

NOTE

1. Charles Thompson has cogently noted that when RIP is computed on fake data, as in the present paper, it becomes the relative index of priority on factitious free recall or RIP-OFF.

(Received for publication January 17, 1974;
revision received March 20, 1974.)