A maximum likelihood nonmetric multidimensional scaling procedure for word sequences obtained in free-recall experiments

KENPEI SHIINA¹

Department of Psychology, School of Letters, Waseda University, Shinjuku-ku, Tokyo 160

A maximum likelihood nonmetric multidimensional scaling procedure is developed for word sequences obtained in free-recall experiments in order to spatially represent the structure of semantic memory. A Monte Carlo simulation showed that this procedure can reproduce given configurations, and a tentative application to empirical data (free-recall data of historical personages by the author as the subject) gave promising results. The procedure has two aspects: mathematical modeling of free-recall and the application of the constant utility model (Luce & Suppes, 1965) to multidimensional scaling. Relations to other maximum likelihood MDS's and the SAM model for memory retrieval (Raaijmakers & Shiffrin, 1981) are discussed.

Key words: free-recall, maximum likelihood, multidimensional scaling, semantic memory, choice axiom, spatial representation.

The sequences of words (or output lists) obtained in free-recall experiments are supposed to reflect the structure of memory. Recently attempts to get spatial representations of words from the word sequences are made by several researchers (Caramazza, Hersh, & Torgerson, 1976; Takano, 1978; Friendly, 1979) in order to study the structure of memory. For example, Friendly proposed several methods which transform order information in word sequences into an inter-word similarity (or distance) matrix: Applying a MDS (Multi-Dimensional Scaling) or a cluster analysis to this matrix yields a graphic representation of words. The transformations adopted in those studies are, however, somewhat ad hoc, and their theoretical foundation is not obvious. This paper introduces a maximum likelihood estimation procedure which directly derives the spatial representation of words from sequences of words. This procedure, which involves the feature of mathematical modeling of free-recall, is essentially one

of "parametric approaches" to the non-metric maximum likelihood MDS (Takane & Carroll, 1981) and it permits statistical model evaluation using the AIC statistic (Akaike, 1974).

The Inner Processes in Free-Recall

It should be mentioned first that our procedure does not cover all types of data arising from so-called free-recall experiments. Its main objective is to reveal the structure of semantic memory (Tulving, 1972). Thus it is restricted to the situation in which words to be remembered are presented randomly to subjects in encoding (Tulving, 1962), in order to avoid non-random aspects of sequences not due to the meaning and interrelationship of words, e.g. serial position effects, seriation, priority of recall for newly remembered words (Murphy & Puff, 1982). These effects are themselves interesting subjects of study, but for the study of "structure of semantic memory" they should be avoided since they are considered as disturbance factors concerning episodic memory (Tulving, 1972). For reasons which we go into in the next section, there is another limitation in this procedure in

¹ The author would like to express his appreciation to Dr. Tatsuro Makino for his encouragement and to Mr. Hiroshi Hojo for his helpful comments.

that it is not always applicable to the incomplete sequences which have unrecalled words, double recalls or extralist intrusions.

We assume that the interrelationship among words can be represented in terms of distances. That is, the words in semantic memory are supposed to be located in K -dimensional Euclidean space (we will often call it semantic space) and the distance d_{ij} between word i and word j is computed by

$$d_{ij} = \left\{ \sum_{l=1}^k (x_{il} - x_{jl})^2 \right\}^{1/2},$$

where x_{il} is the coordinate of word i on axis l . What subjects must do in free-recall experiments can be divided into three parts, namely, encoding, retention, and decoding (retrieval). Although our model is not much concerned with the first two parts, we must make two assumptions about them: First, encoding in free-recall experiments is tagging of the words which are already in the semantic space (Anderson, 1972). Second retention would be perfect when a sufficient number of trials are available to the subjects. When these assumptions are valid, the sequences which subjects produce in the decoding will be more complete as the trials go on, and they are supposed to reflect the structure of seman-

tic memory; the position of a word in a sequence seems to reflect the similarity to the other words in the sequence.

As an example, suppose that in a very simple free-recall experiment, only five animal terms, namely, Lion (L), Tiger (T), Gorilla (G), Orangutan (O), and Dog (D) are presented to a subject, and suppose further that the "true" configuration of the above five words in semantic space is like Fig. 1-a. After a sufficient number of trials, the recalled sequences would become complete. The recalled sequences may be /L/T/D/O/G/, /D/T/L/O/G/, or /O/G/D/T/L/ (this indicates that the items between slashes are recalled words and the recalled order is from left to right). It is probable that the probability of a sequence occurring is not the same across different sequences. For instance, the probability of /L/T/D/O/G/ occurring is greater than that of /L/D/G/T/O/ provided that Fig. 1-a is true, because $P(X|Y)$, the conditional probability that a word X is recalled immediately after a word Y , is expected to be a monotone decreasing function of the distance between X and Y . Thus, $P(T|L) > P(D|L)$, $P(D|T) > P(G|D)$, $P(O|D) > P(T|G)$, so that it seems $P(L/T/D/O/G) > P(L/D/G/T/O)$. This inequality plays an important role in the model to be presented.

The decoding process is conceptualized as a serial searching for the words tagged in the encoding. The searching is just like that of a man in a pitch-dark room trying to find some coins (tagged words) on the floor (semantic space); the only means available for him is to crawl here and there searching for the coins. Here, we can assume that the probability that he finds a coin is a monotone decreasing function of the distance between him and the coin; and after he finds a coin, the probability that he finds another coin is also the decreasing function of the distance between the coin found and the

Fig. 1. True configuration (a) and estimated configurations (b)-(e). The configurations (b)-(e) correspond to the minimum AIC models for S_1 - S_4 shown in Table 1, respectively.

coin to be found. This analogy may be too rough, but it is a probable formulation of the decoding. The word first found among tagged n words in semantic space serves as a reference word for finding the second word among the remaining $n-1$ words, and the first word itself is never searched for thereafter. After the second word is found, it in turn becomes the reference word and is eliminated from the list of the candidates; again the third word is searched for now among the $n-2$ words. Proceeding in this way, we finally obtain a sequence of n words. We have assumed that an unknown function of distance governs this stochastic process. Thus the question now is how distances relate to probabilities.

Luce (1961) argued that $P_S(x; a)$, the probability that stimulus x is chosen as most similar to stimulus a out of a finite set $S(S \ni x, S \not\ni a)$, can be stated as

$$P_S(x; a) = \frac{v(x, a)}{\sum_{y \in S} v(y, a)} \quad (1)$$

where $v(y, a) > 0$ is the scale value (utility) of similarity between y and a on the ratio scale. At the same time functions relating scale values (utilities) to distances have been studied by Shepard (1957) and Schönemann and Wang (1972). This function V must be monotone decreasing and asymptotically approach zero with the increase of distance. We thus consider four alternatives,

$$\begin{aligned} V_1: & v(y, a) = \exp(-c \cdot d_{ya}^2) \\ & \text{(Schönemann \& Wang)} \\ V_2: & v(y, a) = \exp(-c \cdot d_{ya}) \\ & \text{(Shepard)} \\ V_3: & v(y, a) = d_{ya}^{-1} \\ V_4: & v(y, a) = d_{ya}^{-2} \end{aligned} \quad (2)$$

where $c(>0)$ is a constant.² The constant

² Luce (1961) argued V_3 (and also V_4) is unreasonable, since when d_{ya} happens to be 0 it has an infinite scale value. However, so long as y and a are different words, $d_{ya}=0$ is an unrealistic assumption in the present context.

c changes the size but not the shape of the configuration, consequently we set $c=1$ for convenience. Needless to say, we cannot know the real V , but by using the AIC statistic to be discussed later we can choose the V which fits the data best. Given the hypothesized function V , equation (1) becomes a function relating distances to probabilities; it gives the probability that a word x is recalled immediately after a reference word a . In the next section we will consider the likelihood of a sequence using (1).

The Likelihood Function

Let $T = \{1, 2, 3, \dots, n\}$ be a set of n stimulus words, $S(m \times n)$ be the matrix of observed data, where m is the number of observed complete sequences (replications may be over subjects and/or over trials of an individual according to the definition of the hypothesized population), and s_{ij} be the i th sequence's j th recalled word, and $S_{(i)}$ be i th sequence. The probability of $S_{(i)}$ occurring is stated as

$$P(S_{(i)}) = P(s_{i1}/s_{i2}/\dots/s_{in}). \quad (3)$$

It is immediate from the multiplication theorem of conditional probability that equation (3) can be decomposed into products of conditional probabilities;

$$\begin{aligned} & P(s_{i1}/s_{i2}/\dots/s_{in}) \\ &= P(s_{i1})P(s_{i2}/s_{i1})P(s_{i3}/s_{i1}/s_{i2})\dots \\ & P(s_{in}/s_{i1}/s_{i2}/\dots/s_{i(n-1)}). \end{aligned} \quad (4)$$

Assuming that the first word is recalled randomly out of T , we have

$$P(s_{i1}) = 1/n. \quad (5)$$

As noted in the previous section, we assume that free-recall is a successive searching for non-recalled words in semantic space and that the searching process can be approximated by the choice process. Thus, we regard a in (1) as a reference word, x as a word to be recalled, and S as a set of non-recalled words. The important

consequence of this is that (1) is now interpreted as a transition probability from a to x with S (a set of non-recalled words). Then, we relax the conditional probabilities in (4), which imply that $P(s_{ik}/|s_{i1}/s_{i2}/\dots/s_{i(k-1)})$ is strictly dependent on the recall history $|s_{i1}/s_{i2}/\dots/s_{i(k-1)}|$, into the transition probabilities;

$$\begin{aligned} P(s_{i2}/|s_{i1}/) &= P_{T_1}(s_{i2}; s_{i1}) \\ P(s_{i3}/|s_{i1}/s_{i2}/) &= P_{T_2}(s_{i3}; s_{i2}) \\ &\vdots \\ P(s_{in}/|s_{i1}/s_{i2}/\dots/s_{i(n-1)}) &= P_{T_{n-1}}(s_{in}; s_{i(n-1)}) \end{aligned} \quad (6)$$

where $T_1, T_2, \dots, T_{n-1}$ (the sets of non-recalled words) are $T_1 - \{s_{i1}\}, T_2 - \{s_{i2}\}, \dots, T_{n-2} - \{s_{i(n-1)}\}$, respectively. This relaxation amounts to the assumption that k th recall is not strictly dependent on the recall history $|s_{i1}/s_{i2}/\dots/s_{i(k-2)}|$. That is, while the right-side hand of (6) is still a function of the set of non-recalled words and the reference word, they are independent of which specific permutation of $\{s_{i1}, s_{i2}, \dots, s_{i(k-2)}\}$ actually occurred. Assuming (1), we can rewrite the right-hand side of (6) as

$$\begin{aligned} P_{T_1}(s_{i2}; s_{i1}) &= v(s_{i2}, s_{i1}) / \sum_{y \in T_1} v(y, s_{i1}) \\ P_{T_2}(s_{i3}; s_{i2}) &= v(s_{i3}, s_{i2}) / \sum_{y \in T_2} v(y, s_{i2}) \\ &\vdots \\ P_{T_{n-1}}(s_{in}; s_{i(n-1)}) &= v(s_{in}, s_{i(n-1)}) / \sum_{y \in T_{n-1}} v(y, s_{i(n-1)}). \end{aligned}$$

so that the likelihood of $S_{[i]}$ becomes

$$\begin{aligned} P(S_{[i]}) &= \frac{1}{n} \prod_{k=2}^n \{v(s_{ik}, s_{i(k-1)}) / \sum_{y \in T_{k-1}} v(y, s_{i(k-1)})\}. \end{aligned} \quad (7)$$

The likelihood function Q for the entire data set is given by

$$Q = \prod_{i=1}^m P(S_{[i]}). \quad (8)$$

Accordingly, by maximizing

$$\ln(Q) = \sum_{i=1}^m \ln \{P(S_{[i]})\}, \quad (9)$$

with respect to the coordinates of the n words, we obtain maximum likelihood estimates of the coordinates of the words.

Independence of the transition probabilities. We have assumed the independence of $P_{T_{k-1}}(s_{ik}; s_{i(k-1)})$ and $P_{T_{k-1}}(s_{ik'}; s_{i(k'-1)})(k \neq k')$ in (7), and the independence of $P(S_{[k]})$ and $P(S_{[k']})(k \neq k')$ in (8). Independence of observations is necessary in order to obtain the maximum likelihood estimates. Thus, we can hardly analyse such data that violate this condition. When $S_{[i]}$'s are taken from different individuals, the independence condition in (8) will be satisfied, while in such a case where $S_{[i]}$'s include an individual's replications, some experimental treatments which ensure the independence will be necessary.

The independence condition in (7) is, however, somewhat problematic. The independence of $P_{T_{k-1}}(s_{ik}; s_{i(k-1)})$ and $P_{T_{k'-1}}(s_{ik'}; s_{i(k'-1)})(k \neq k')$ implies that a transition from $s_{i(k-1)}$ to s_{ik} with T_{k-1} and a transition from $s_{i(k'-1)}$ to $s_{ik'}$ with $T_{k'-1}(k \neq k')$ are statistically independent. Although this assumption seems to be too strong, it would be sufficient as a first approximation.

Incomplete sequences. We have so far been concerned with the likelihood function for complete sequences. The likelihood function (7), however, does not exclude incomplete sequences which have unrecalled words, while we have no provision for the sequences which have double recalls or extralist intrusions. That is, if we modify the n in (7) to n_i (n_i refers to the number of recalled words of $S_{[i]}$), we can deal with, in theory, incomplete sequences. Nevertheless, a serious problem arises when we analyse such sequences. Remember that our model is based upon the two assumptions that, in an appropriate experimental condition, the tagging of words can be complete and the effects concerning episodic memory

is negligible. Generally, the assumptions do not hold for incomplete sequences, so that we need some theoretical and/or empirical justification in dealing with them. Hence, when (7) is modified to accept incomplete sequences, care must be taken not to include inappropriate incomplete sequences.

Relation to the multinomial distribution. It is noteworthy that, when we consider only the complete sequences, (9) can be regarded as the log-likelihood of a reparametrized multinomial distribution. Observe that the number of sequences which theoretically can occur is $n!$. Without loss of generality, we can denote by $\xi_j (j=1, 2, \dots, n!)$ the events that these $n!$ sequences occurring. Further, let the number of observations be l and let l_j be the number of ξ_j occurring.

The log-likelihood of the multinomial distribution is given by

$$\sum_{j=1}^{n!} l_j \ln \{P(\xi_j)\} + \text{constant}, \quad (10)$$

where
$$\sum_{j=1}^{n!} P(\xi_j) = 1.$$

If we substitute $P(\xi_j)$'s in (10) by corresponding $P(S_{(i)})$'s, l_j 's by the number of the occurrences of corresponding $S_{(i)}$'s, and ignore the constant term, then (10) becomes equal to (9). This means that the likelihood function (9) is nothing but the log-likelihood of a multinomial distribution reparametrized by the coordinates of stimulus words. Note that once we obtain the estimates of those coordinates of words, we have non-zero estimates of $P(\xi_j)$'s by (7) even when n is large, it is impossible to have sufficiently larger observations than $n!$. The question arises as to whether $\hat{P}(\xi_j)$'s estimated from (9) always satisfy the requirement:

$$\sum_{j=1}^{n!} \hat{P}(\xi_j) = 1.$$

This can be easily proved as shown in

Appendix 1.

The maximization procedure. In order to maximize $\ln(Q)$, we need the first and second derivatives of $\ln(Q)$ with respect to the unknown coordinates. The analytical partial differentiation of $\ln(Q)$ is possible but too complex, so that difference approximations of derivatives are used (see Appendix 2). The inverse matrix of the second derivatives (Hessian) is also approximated by the BFGS method (Dennis & Moré, 1977) using the above partial derivatives.

Model selection. The AIC statistic (Akaike, 1974), which is defined by $AIC = -2 \times (\text{maximum log likelihood}) + 2 \times (\text{effective number of parameters})$, is used to choose the best fitting model with respect to the dimensionality and the given function V . The effective number of parameters is given by $nk - k(k-1)/2 - k$ for V_1 and V_2 and $nk - k(k-1)/2 - k - 1$ for V_3 and V_4 , where n is the number of stimulus words and k is the number of dimensions, since the translational and rotational invariance holds for the solution for V_1 - V_4 and the indeterminacy of the unit for V_3 and V_4 . It is well-known that the AIC is a badness-of-fit index. Thus, as the model fits the data better, the AIC becomes smaller. As for examples of studies in which this statistic was used effectively, see Takane (1981), Takane and Carroll (1981), Takane and Sergent (1983), and Hojo (1982, 1983). Unfortunately, the AIC is applicable only to large sample problems. Thus, in the examples to be presented in the next section we restrict, following Takane (1981), the number of parameters to be estimated so as not to exceed one-tenth of the number of observations.

Examples

Artificial Data (Monte Carlo Simulation)

In order to test the reproducibility of the procedure, four sets (S_1, S_2, S_3, S_4) of 100 artificial sequences were randomly generated following the equations (1), (2),

Table 1
AIC values for 48 solutions

Hypothesized function	Dimensionality	Data sets			
		S_1	S_2	S_3	S_4
V_1	1 $df=4$	842.19	553.46	917.77	836.50
	2 $df=7$	835.89 ^{a)}	465.73	919.04	841.90
	3 $df=9$	839.89	469.06	923.04	840.29
V_2	1 $df=4$	842.13	616.38	921.69	848.12
	2 $df=7$	836.13	465.58 ^{a)}	919.12	838.29
	3 $df=9$	840.13	469.58	923.12	842.29
V_3	1 $df=3$	842.39	605.21	924.42	836.18
	2 $df=6$	846.33	611.15	917.20 ^{a)}	837.94
	3 $df=8$	849.84	615.15	921.20	841.94
V_4	1 $df=3$	868.35	610.07	998.16	866.18
	2 $df=6$	836.00	574.18	940.24	834.46 ^{a)}
	3 $df=8$	840.00	578.18	944.24	838.30

^{a)} The minimum AIC solution for each data set.

(4), and (5) using Fig. 1-a as the true configuration. The four sets S_1 , S_2 , S_3 , S_4 correspond to the choice among the functions V_1 , V_2 , V_3 , V_4 , respectively. Then the four sets were used to reproduce the "true" configuration shown in Fig. 1-a, using the four functions as hypothesized functions and varying the dimensionality from one to three. Therefore 48 solutions were obtained in total. Figure 1b-e show the configurations represented by means of the best fitting models for the data sets S_1 , S_2 , S_3 , S_4 , respectively (minimum AIC models) and Table 1 shows the AIC values

for the corresponding 48 solutions. As expected, Table 1 depicts that the "true" models always give the minimum AIC values and the configurations correspond to the minimum AIC models are similar to that of Fig. 1. These results demonstrate the feasibility of our procedure and the usefulness of the AIC statistic for model selection.

Empirical Data

A simple multi-trial free-recall experiment was carried out to obtain relevant data for our procedure. Stimulus words

Fig. 2. Three dimensional configuration of 12 words (the minimum AIC model shown in Table 2).

Table 2
AIC values for eight solutions

Dimensionality	Hypothesized function			
	V_1	V_2	V_3	V_4
2	1 910.70 ($df=21$)	1 952.42 ($df=21$)	2 074.86 ($df=20$)	1 918.89 ($df=20$)
3	1 853.64 ^{a)} ($df=30$)	1 865.80 ($df=30$)	2 012.36 ($df=29$)	1 911.56 ($df=29$)

^{a)} The minimum AIC solution.

were names of 12 famous historical Europeans shown in Fig. 2 and the subject was the author.

Encoding phase. For each trial the 12 words were successively presented on a video display screen. Of course the order of presentation was randomly changed from trial to trial, and the inter-stimulation interval was approximately 1 s. At the end of each trial he recalled as many the words as he could. The intertrial interval was 30 s and all recalls (spoken) were tape-recorded.

Recall phase. When recalled sequences became complete, the presentation of words was stopped. From that time, the subject tried to recall the 12 words as fast as and as many as possible with the intertrial intervals ranging from 3 min to 6 h (median 5 min). During the intertrial intervals the subject did his daily business (this probably increased the independence of observations). We obtained 60 complete sequences out of a total of 78 sequences; all the complete sequences were used for the parameter estimation.

Results. Table 2 shows the AIC values across dimensionality and the four functions. The best fitting model determined by means of the AIC was V_1 's 3-dimensional solution. Figure 2 shows the configuration of the words by means of this best fitting model. Unfortunately, we can hardly interpret the axes clearly, but there are several overlapping meaningful clusters. We can make at least the following statements: (1) Plato and Caesar are

located apart from the other persons possibly because they are ancient, (2) we find the clusters of nationalities (French, German, and English), (3) further, we can see the clusters of professions (politicians and generals, artists, and scientists). Figure 3 shows the relation between the estimated distances and the mean interitem response times. It can be expected that the larger the interword distance is, the longer the interitem response time becomes, for reason that a transition from a word to another is naturally assumed to be a time-consuming process. Figure 3

Fig. 3. The relation between estimated distance and inter-recall response time.

shows that, roughly speaking, this holds true for the present data (the correlation coefficient is .45) and this result gives additional validity to our procedure as modeling of internal processing of free-recall.

Discussion

We briefly discuss two aspects of our procedure: one of nonmetric maximum likelihood MDS, and the other of memory model.

The relation to other MDS models. The nonmetric maximum likelihood MDS's of Zinnes and Griggs (1974), Takane (1981), Takane and Carroll (1981), Takane and Sergent (1983), and Hojo (1982, 1983) assume that d_{ij} is error-perturbed and generates an error-perturbed metric process λ_{ij} such as

$$\lambda_{ij} = \varepsilon(d_{ij}, e_{ij}), \quad (11)$$

where e_{ij} is an error random variable and ε is a joint function of d_{ij} and e_{ij} (Takane

& Carroll, 1981). The likelihood functions are defined by assuming appropriate response models which reduce metric process λ into nonmetric data. In contrast our model and the model proposed by Schönemann and Wang (1972) do not assume such an error-perturbed metric process explicitly; instead they assume the functions which directly relate distances to utilities.

The difference between the above two kinds of models is best explained by considering the former as the random (or Thurstonian) utility model and the latter as the constant (or strict) utility model (Luce & Suppes, 1965). The relationship between these two kinds of models was studied by Yellott (1977); he proved that only when the random metric processes (or discriminial processes) are double-exponentially distributed, the random utility model is equivalent to the strict utility model provided that the random variables differ only in their means and the choice set has more than three alternatives. Consequently, if we assume

$$Ps(x, a) = P[-d_{xa} + e_{xa} = \text{Max} \{-d_{ya} + e_{ya} \mid S \ni y\}]$$

or

$$Ps(x, a) = P[-d_{xa}^2 + e_{xa} = \text{Max} \{-d_{ya}^2 + e_{ya} \mid S \ni y\}],$$

where e_{xa} and e_{ya} are independent identical double-exponentially distributed variables, we have (1) and the functions $V_1: v(x, a) = \exp(-d_{xa})$ and $V_2: v(x, a) = \exp(-d_{xa}^2)$, respectively (see Yellott, Lemma 6). In this sense the model presented in this paper can be regarded as one of the Thurstonian models. Needless to say, however, it is possible to assume the constant utility model so that V_3 and V_4 continue to be possible alternatives.

The relation to other memory models. The scale value $v(x, a)$ may be regarded as the strength of an association between a and x or the level of activation spreading from a to x (Collins & Loftus, 1975). Then the functions in (2) can be considered as the ones representing associa-

tion strength in terms of distance. In this sense our model can be considered as one of the network theories of semantic memory. Among others we consider the relation to the SAM (Search of Associative Memory) model (Gillund & Shiffrin, 1984; Raaijmakers & Shiffrin, 1981). The SAM model for simple free-recall version assumes that the probability of sampling an image³ W_i from long-term memory is

³ The usage of the term "image" in the SAM model is special. The SAM model for free-recall assumes that the memory search consists of a sampling phase, in which information concerning an item called "image" is sampled from long-term memory, and a recovery phase, in which the name of the sampled "image" is recovered according to a distributional function.

given by

$$P(W_i|C) = S(C, W_i) / \sum_{j=1}^n S(C, W_j),^4 \quad (12)$$

or

$$P(W_i|C, W_k) = \frac{S(C, W_i)S(W_k, W_i)}{\sum_{j=1}^n S(C, W_j)S(W_k, W_j)} \quad (13)$$

where C denotes the general temporal-contextual cue specific to the experiment, W_k , item k as a retrieval cue, n , the number of images, and $S(x, y)$, the strength with which the cue x samples the image y . The SAM model further assumes that $S(x, y)$ is determined by three factors: (a) rehearsal and coding processes, (b) pre-existing associations, and (c) match of the cue encoding at study and test (the encoding specificity principle, Tulving and Thompson, 1973), and that the contribution of factor (a) to $S(x, y)$ is proportional to the total amount of time x and y are simultaneously in the rehearsal buffer. Equation (12) is assumed when the context cue C alone is used in retrieval, and (13) is assumed when both context cue C and item W_k are used.

We show that in an experimental condition as supposed in this paper where to-be-remembered items are randomly presented many times, our model is very similar to the SAM model. When learning trials are repeated many times, $S(C, W_i)$, the strength between the general context cue and an image W_i , will be the same value for all W_i 's, because the probability of C and W_i simultaneously in the rehearsal buffer will be equal for all W_i 's provided that W_i 's are randomly presented as supposed in our model. Then we have $(12) = 1/n$ and $S(C, W_i)$'s in (13) are cancelled out. Note that this yields virtually the same equations as (1)

and (5). Further, the total amount of time items x and y are simultaneously in the rehearsal buffer will also asymptotically be the same for all the pairs of x and $y(x, y \in T)$, so that factor (a) will be constant. The factor (c) will also be constant when the meanings of the words do not change at study and test. This means that two of the three factors in $S(x, y)$ will be constant, so that only factor (b) remains. Therefore now $S(x, y)$ becomes the index of the pre-existing association between x and y , which is similar to $v(x, y)$. This result gives us a basis to consider that the experimental condition assumed to be necessary for our model is not groundless.

Future directions. Many researchers have proposed mathematical models of memory. Those models often lack parameter estimation techniques and the criterion that enables us to decide which model is better. Although our procedure is applicable only to very limited data, it is able to shed light on internal processings and choose the best fitting model among possible models. Of course our model has large room for refinement; for instance, the representational model of memory structure need not to be spatial. Possibly, Tversky's feature matching model (1977) or the EBA model (1972) may also be considered. Nevertheless, the fact that this procedure can select the best fitting model through the use of the AIC statistic is a prominent feature, so that the author believes that this kind of approach will contribute much to the study of memory. At any rate, further studies are in order using this procedure so as to demonstrate the validity of the model and the effectiveness of the procedure.

References

- Akaike, H. 1974 A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, **19**, 716-723.
- Anderson, J. R. 1972 FRAN: A simulation model of free-recall. In G. H. Bower (Ed.), *The*

⁴ Redundant subscripts in original notation are omitted.

- psychology of learning and motivation*. Vol. 5. New York: Academic Press. Pp. 315-378.
- Caramazza, A., Hersh, H., & Torgerson, W. S. 1976 Subjective structures and operations in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, **15**, 103-117.
- Collins, A. M., & Loftus, E. F. 1975 A spreading-activation theory of semantic processing. *Psychological Review*, **82**, 407-428.
- Dennis, J. E., & Moré, J. J. 1977 Quasi-Newton method, motivation and theory. *SIAM Review*, **19**, 46-89.
- Friendly, M. 1979 Methods for finding graphic representations of associative memory structures. In C. R. Puff (Ed.), *Memory organization and structure*. New York: Academic Press. Pp. 85-129.
- Gillund, G., & Shiffrin, R. M. 1984 A retrieval model for both recognition and recall. *Psychological Review*, **91**, 1-67.
- Hojo, H. 1982 Confusion-choice model for multidimensional scaling: Modification of Nakatani's model. *Japanese Psychological Research*, **24**, 211-215.
- Hojo, H. 1983 A maximum likelihood method for additive clustering and its applications. *Japanese Psychological Research*, **25**, 191-201.
- Luce, R. D. 1961 A choice theory analysis of similarity judgments. *Psychometrika*, **26**, 151-163.
- Luce, R. D., & Suppes, P. 1965 Preference, utility, and subjective probability. In R. D. Luce, R. R. Bush & E. Galanter (Eds.), *Handbook of mathematical psychology*. Vol. 3. New York: Wiley. Pp. 249-410.
- Murphy, M. D., & Puff, C. R. 1982 Free-recall: Basic methodology and analysis. In C. R. Puff (Ed.), *Handbook of research methods in human memory and cognition*. New York: Academic Press. Pp. 99-128.
- Raaijmakers, J. G. W., & Shiffrin, R. M. 1981 Search of associative memory. *Psychological Review*, **88**, 93-134.
- Schönemann, P. H., & Wang, M. M. 1972 An individual difference model for the multidimensional analysis of preference data. *Psychometrika*, **37**, 275-309.
- Shepard, R. N. 1957 Stimulus and response generalization: a stochastic model relating generalization to distance in psychological space. *Psychometrika*, **22**, 325-345.
- Takane, Y. 1981 Multidimensional successive categories scaling: a maximum likelihood method. *Psychometrika*, **46**, 9-28.
- Takane, Y., & Carroll, J. D. 1981 Nonmetric maximum likelihood multidimensional scaling from directional rankings of similarities. *Psychometrika*, **46**, 389-405.
- Takane, Y., & Sergent, J. 1983 Multidimensional scaling models for reaction times and same-different judgments. *Psychometrika*, **48**, 393-423.
- Takano, Y. 1978 A dissimilarity analysis for organization in multitrial cued retrieval. *Japanese Psychological Research*, **20**, 148-153.
- Tulving, E. 1962 Subjective organization in free-recall of "unrelated" words. *Psychological Review*, **69**, 344-354.
- Tulving, E. 1972 Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory*. New York: Academic Press. Pp. 381-403.
- Tulving, E., & Thompson, D. M. 1973 Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, **80**, 352-373.
- Tversky, A. 1972 Choice by elimination. *Journal of Mathematical Psychology*, **9**, 341-367.
- Tversky, A. 1977 Features of similarity. *Psychological Review*, **84**, 327-352.
- Yellott, J. I., Jr. 1977 The relationship between Luce's choice axiom, Thurstone's theory of comparative judgement, and the double exponential distribution. *Journal of Mathematical Psychology*, **15**, 109-144.
- Zinnes, J. L., & Griggs, R. A. 1974 Probabilistic, multi-dimensional unfolding analysis. *Psychometrika*, **39**, 327-350.

(Received Dec. 3, 1984; accepted
Jan. 18, 1986)

Appendix 1

For $\xi = |S_1/S_2/\dots/S_n|$ which is an arbitrary member of $\xi_j (j=1, 2, \dots, n!)$, let $T_1, T_2, \dots, T_{n-1}$ be the sets of words such that $T_1 = T - \{S_1\}$, $T_2 = T_1 - \{S_2\}$, ..., $T_{n-1} = T_{n-2} - \{S_{n-1}\}$, respectively. Note that T_k is determined only when $|S_1/S_2/\dots/S_k|$ is determined. Observe that

$$\begin{aligned} \sum_{j=1}^{n!} \hat{P}(\xi_j) &= \sum_T 1/n \left[\sum_{S_2 \in T_1} \{v(S_2, S_1) / \sum_{y \in T_1} v(y, S_1)\} \right. \\ &\quad \times \left[\sum_{S_3 \in T_2} \{v(S_3, S_2) / \sum_{y \in T_2} v(y, S_2)\} \right. \\ &\quad \vdots \\ &\quad \times \left[\sum_{S_{n-2} \in T_{n-3}} \{v(S_{n-2}, S_{n-3}) / \sum_{y \in T_{n-3}} v(y, S_{n-3})\} \right. \\ &\quad \times \left. \left. \left[\sum_{S_{n-1} \in T_{n-2}} \{v(S_{n-1}, S_{n-2}) / \sum_{y \in T_{n-2}} v(y, S_{n-2})\} \right] \dots \right] \right]. \end{aligned}$$

Obviously,

$$\sum_{S_{k-1} \in T_{k-1}} \{v(S_{k-1}, S_{k-2}) / \sum_{y \in T_{k-1}} v(y, S_{k-2})\} = 1$$

$$(k = 2, 3, \dots, n)$$

so that

$$\sum_{j=1}^{n+1} \hat{P}(\xi_j) = \sum_T 1/n = \sum_{k=1}^n 1/n = 1.$$

Appendix 2

The derivatives are approximated by

$$\frac{\partial \ln(Q)}{\partial x_{il}} \doteq \frac{\ln\{Q(x_{il}+h)\}-\ln\{Q(x_{il})\}}{h}$$

or

$$\frac{\partial \ln(Q)}{\partial x_{il}} \doteq \frac{\ln\{Q(x_{il}+h)\}-\ln\{Q(x_{il}-h)\}}{2h}$$

where h is the step length ($h=.0001$).