JOURNAL OF MATHEMATICAL PSYCHOLOGY 13, 214–240 (1976)

Verification of Algebra Step Problems: A Chronometric Study of Human Problem Solving*

PAUL G. MATTHEWS AND RICHARD C. ATKINSON

Stanford University, Stanford, California 94305

A class of simple problem solving tasks requiring fast accurate solutions is introduced. In an experiment subjects memorized a mapping rule represented by lists of words labeled by cue words and made true/false decisions about conjunctions of propositions of the form "Y is in the list labeled by X," written "$X \rightarrow\!\!\!\!- Y$". Response times are analyzed using a "stage modeling" technique where problem solving algorithms are composed using a small set of psychological operations that have real time characteristics specified parametrically. The theoretical analysis shows that response time performance is adequately described in terms of the sequential application of elementary psychological operations. Unexpectedly, it was found that the proposition "$X \rightarrow\!\!\!\!- Y$ and $X \rightarrow\!\!\!\!- Z$" was verified as quickly as the apparently simpler "$X \rightarrow\!\!\!\!- Y$". A case is presented for the modeling technique as applied to memory and problem solving tasks in terms of theoretical parsimony, statistical simplicity, and flexibility in investigative empirical research. Suggestions are made as to possible theoretical relations among fast problem solving, more complex and slower problem solving, and research in fundamental memory processes.

INTRODUCTION

The dominant theoretical approach to the analysis of problem solving has been to construct a formal model, often in the form of a computer program, that simulates some qualitative aspects of human problem solving performance such as the protocol sequences observed in deriving logic theorems (Newell & Simon, 1972). In these analyses emphasis is placed on the integration of elementary information operations into a problem solving algorithm while less attention is given to the elementary operations themselves. An approach that has been relatively less well explored is to specify the processing time implications of proposed algorithms and to determine whether observed human response times (RT's) are consistent with the predicted pattern. From a statistical point of view, problems that require several minutes to solve or involve extensive searching for a solution (e.g., looking for the best move in a chess

* This research was supported by National Science Foundation Grant NSF-GJ-443X3, National Institute of Mental Health Grant MH21747 and a National Research Council of Canada Postgraduate Scholarship to the first author.

position, de Groot, 1965) might be expected to have large RT variances even for an individual subject such that it becomes impractical to model the fine details of RT. However, for simple problems where human subjects are easily able to respond correctly in a matter of a few seconds, it should be possible to verify the processing time predictions of specific problem solving algorithms.

One method for deriving RT predictions is to describe problem solving algorithms in terms of the sequential application of a set of basic psychological operations (procedures, subroutines, or "stages") each of which requires real processing time and has some probability of producing an error. Leaving the details for later discussion, the theoretical RT for an algorithm applied to a particular problem can be described as the sum of the processing times of the operations applied and the error rate is roughly one minus the product of the correct probabilities of these operations. An alternative technique for making RT predictions is to assign computational complexity measures to the basic operations and to derive the complexity of an algorithm as the sum of the complexities of its component operations; computational complexity is then directly interpreted as linearly related to theoretical mean RT. This complexity assignment method yields the same description of mean RT's as does the corresponding stage model although it does not describe higher RT moments. Note that both methods are easily generalized to take account of the possibility of mixed (randomized) strategies for applying available algorithms.

On a general theoretical level, the RT analysis of fast accurate problem solving can be a valuable source of evidence in deciding on a set of basic psychological operations used in human problem solving. The case is similar to that for chronometric studies of linguistic comprehension (Chase & Clark, 1972), where alternative representations of propositions can sometimes be discriminated by constructing RT models for processing propositions to make true/false decisions. For problem solving theories it is desirable to build algorithms working with a set of elementary operations which have some preferred characteristics, such as corresponding to procedures or subroutines that can be conveniently written as logical units when programming in a particular language, or being general in the sense that the same set of operations can be used in solving several types of problems. Another preferred characteristic is that the set of operations has "psychological validity" insofar as real time processing aspects of the operations can be defined and verified in observed RT performance.

Algebra Step Problems

To pursue these ideas an experimental task was sought where subjects would learn a set of rules (e.g., the moves of pieces in a board game, or a mapping of one set of objects into another) and be required to solve true/false problems by repeated application of these rules. It was thought that a model for the single application of a rule

could then be extended to a model for the entire problem solving task by specifying the way rules were applied to solve a problem.

Consider a small finite set X and a rule that assigns to each element of X a subset of X. Such a rule can be written in the form of a *transition table* such as that in Fig. 1

x_1—x_7
x_2—x_8
x_3—x_9
x_4—x_1 , x_5
x_5—x_6 , x_2
x_6—x_3 , x_4
x_7—x_2 , x_4 , x_8
x_8—x_9 , x_3 , x_5
x_9—x_6 , x_7 , x_1

FIG. 1. ASP transition table.

which was used in an experiment to be described later. A memorized transition table, say where X is a set of consonant-vowel-consonant (CVC words, might be represented as "lists" in some memory store with "addresses" corresponding to the elements of X. One of the most basic propositions that can be made about a particular transition table is that x_i is mapped into a list that contains x_j , written $x_i \rightarrow\!\!-\ x_j$ as a ***mapping diagram***, where x_i and x_j are variables standing for elements of X; this proposition is either true or false. A subject who has memorised a transition table can be presented with the proposition $x_i \rightarrow\!\!-\ x_j$ and be required to make a true/false decision using his knowledge of the rule as defined by the table. In the experiment to be described, subjects were presented with logical "and" conjunctions of these simple propositions and RT's for a true/false decision were measured. The propositional forms or ***problem types*** used are listed in Fig. 2 in three groups (A, B and C) according to the geometric shapes of the mapping diagrams. A problem is true if and only if all the propositions represented by the arrows or ***links*** are true; if just one link is false then the problem is false. For example, P(——) in Fig. 2 is true only if $x_i \rightarrow\!\!-\ x_j$ and $x_j \rightarrow\!\!-\ x_k$ and $x_k \rightarrow\!\!-\ x_l$; it is false if any one of these propositions is false. Similarly, P(−<) is true only if $x_i \rightarrow\!\!-\ x_j$ and $x_j \rightarrow\!\!-\ x_k$ and $x_j \rightarrow\!\!-\ x_l$; and P(>−) only if $x_i \rightarrow\!\!-\ x_k$ and $x_j \rightarrow\!\!-\ x_k$ and $x_k \rightarrow\!\!-\ x_l$.

In the experiment subjects memorized transition tables of the form represented in Fig. 1 where the elements of X were CVC words, and were tested with problems of the sort illustrated in Fig. 2. Representing a transition table in memory as stored lists, an individual link, $x_i \rightarrow\!\!-\ x_j$, could be verified true or false by using the *cue* x_i to "access" the appropriate list in memory and then "scanning" the *probe* x_j against this list for a "match"; if a match is obtained then the link is true and otherwise false. A model for the verification of the conjunctive propositions could then be obtained on the assumption that verification proceeds one link at a time in some specified order. These notions are developed in the discussion section below. Since it is possible to verify

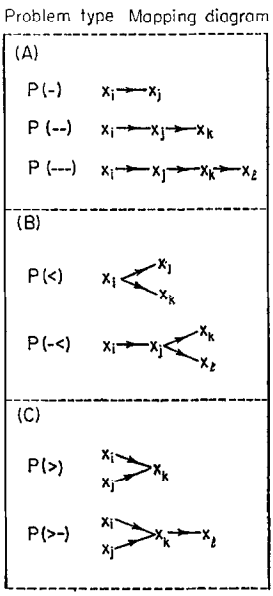


FIG. 2. ASP problem types.

mapping diagrams by checking each link in a step by step manner, the test items used in this task are referred to as *algebra step problems* (ASP).

ASP items for each problem type were selected from a computer generated listing of all possible items given the transition table (Fig. 1) such that within each problem type there were an equal number of true and false items. For false problems exactly one link was false, and for each type the false link occurred with equal frequency at each link position. In addition an effort was made to match the frequencies of occurrence of CVC words between true and false items within each problem type so as to avoid a possible source of response bias. The total pool of about 400 distinct items was divided into four blocks; the same P(–) items occurning in each block but otherwise there was no overlap. Denoting blocks by B_1, B_2, B_3, B_4, subjects were tested over six sessions with one block per day in the order $B_1B_2B_3B_4B_1B_2$, where blocks and trials within blocks were randomised for each subject individually. A set of nine CVC words was randomly assigned to the abstract transition table scheme for each subject: no two subjects had identical transition tables although all tables had the same formal structure.

The experiment was run using an Imlac Corporation PDS-1 cathode ray tube (CRT) display and keyboard, interfaced with a PDP-10 computer. Six female subjects ran for seven sessions; the first session was devoted to a transition table learning drill (subjects did not memorize their transition tables prior to the first session), and the remaining six sessions were used for ASP test items. On a single trial of the drill a cue word was

presented on the CRT and the subject was required to type the appropriate list in serial order (since the CVC words used had unique initial consonants, the subject typed only the first letter of each word and the computer completed the words with suitable horizontal spacing). On completing her response to a cue the subject pressed the keyboard spacebar and the correct list was printed horizontally directly beneath the typed response, providing feedback and an opportunity for study. Permutations of the nine cue words were run and following each permutation the subject was told her percentage of correct responses and the time taken to respond to all the cues. Subjects were required to participate in the drill until they could consistently achieve perfect accuracy with a response time under 25 sec; all subjects met this criterion within 30–50 min of the drill.

In the problem sessions ASP items were displayed at the center of the CRT and subjects responded true/false using two keys on the lower row of the keyboard. The subject initiated trials by pressing the spacebar following a ready signal. Items were preceded by a 1 sec duration fixation cross and appeared just to the right of the cross, remaining on the screen until the subject responded. Immediately after responding subjects received a feedback message indicating correctness and response time.

Before the first problem session subjects were shown examples of the seven problem types and told to respond "true" if and only if all the links in an item were true and to respond "false" as soon as they knew that one link was false. Subjects were informed that there were an equal number of true and false items within each problem type on each day and that false items had exactly one false link which was equally likely to occur in any position. On the first day of problems subjects were instructed to be completely accurate for the initial 30 or 40 trials and then to increase their speed as they got a feeling for the task. For subsequent testing sessions subjects were instructed to respond as quickly they could without making more than about one error in 20 trials on average. Subjects were explicitly instructed never to guess and never to "think twice" about their response once they had made a decision.

Experimental Results

To eliminate early practice effects and to facilitate the observation of stable task strategies the data for each subject from the first of the six testing sessions was discarded together with the first ten trials of the remaining five sessions, yielding on the order of 550 trials per subject. Only correct RT's excluding outliers were analysed. Correct RT histograms were plotted separately for each problem type, for both true and false responses, and for each subject to identify possible outliers. Response times falling more than 1 sec above the main distribution as determined by the mode and the contiguous tails were eliminated; such outliers constituted about 2% of the correct RT data.

Due to the complex description of the ASP items it is not possible to represent all aspects of the data simultaneously in a single graph or table. However, by collapsing across various subsets of the data we can obtain a reasonable picture of major effects which can then direct more detailed modeling and statistical evaluation. Since plots of data for individual subjects showed subjects to be qualitatively comparable, the RT data for all six subjects were pooled to simplify the presentation of results. Table I presents RT and error rate data classified by problem type and position of the false link (if any). The notation P(---) TTF indicates that the third link from the left was false; P(>–) FTT that the upper link of the branch (>) was false; P(–<) TTF that the lower link of the branch (<) was false. Observed means and variances and theoretical means (derived from a statistical model introduced below) are averages across subjects weighted by the numbers of correct RT's observed.

TABLE I

Group RT Means and Errors[a]

Type	False link	Obs mean (msec)	Th mean (msec)	Obs SD (msec)	Error (%)	Total (*N*)
P(–)	T	1576	1529	590	5.6	245
P(–)	F	2041	1993	741	5.3	243
P(––)	TT	2468	2541	905	2.3	251
P(––)	FT	2101	1842	862	7.8	117
P(––)	TF	3161	3187	1035	3.8	121
P(–––)	TTT	3631	3584	1103	4.3	388
P(–––)	FTT	2137	1871	949	7.5	121
P(–––)	TFT	3374	3086	1252	6.7	125
P(–––)	TTF	4184	4046	1106	10.5	121
P(<)	TT	1580	1567	532	4.5	161
P(<)	FT	2287	2068	815	8.3	75
P(<)	TF	2152	2014	964	3.0	88
P(–<)	TTT	2592	2534	933	4.3	328
P(–<)	FTT	1950	1825	994	10.2	112
P(–<)	TFT	2972	3001	895	5.9	108
P(–<)	TTF	2864	3000	858	6.8	112
P(>)	TT	2575	2610	752	5.1	121
P(>)	FT	2501	2211	835	5.0	63
P(>)	TF	3002	2876	735	3.1	61
P(>–)	TTT	3657	3539	1045	3.7	237
P(>–)	FTT	2622	2406	1146	4.2	78
P(>–)	TFT	3342	2990	1169	6.7	79
P(>–)	TTF	4002	4153	1060	1.7	73

[a] By problem type and position of false link.

Figures 3 and 4 are based on the data of Table I; curves represent theoretical mean RT. Figure 3 plots true and false mean RT by problem types. A striking feature about these data are the following approximate equalities of mean RT's obtaining among the problem types:

$$P(<) = P(-) \quad \text{and} \quad P(-<) = P(--)$$
$$P(>) = P(--) \quad \text{and} \quad P(>-) = P(---).$$

Of course these equalities hold among averages including quite distinct items within each problem type, but they do suggest that the time to verify a left branch configuration ($<$) is not substantially different from the time for a simple link ($-$). In contrast, verifying two links in the ($>$) configuration appears to take the same time as two links

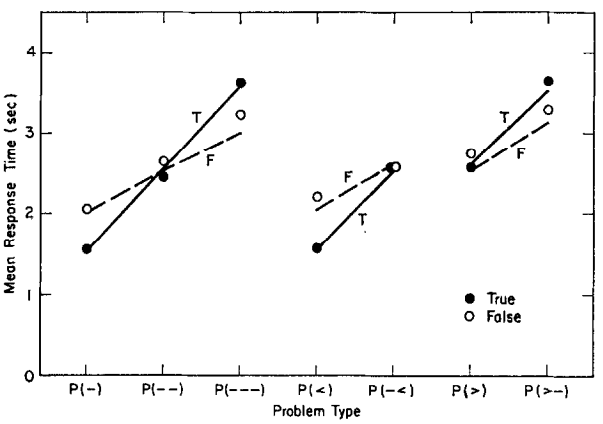


FIG. 3. Mean correct RT's plotted by problem type and true/false. (Points are data and curves are theoretical.)

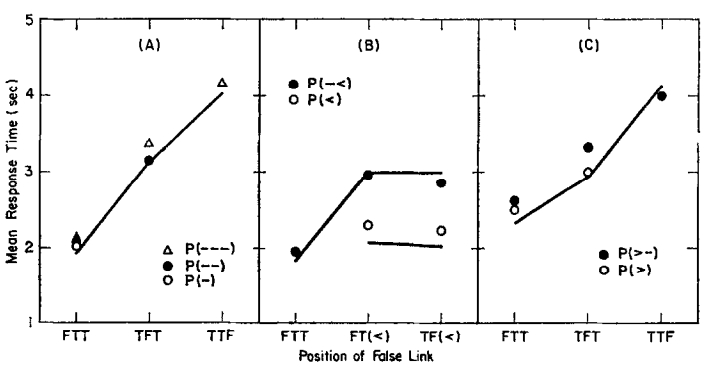


FIG. 4. Mean correct RT's for false items plotted by problem type and position of false link.

in the (—) configuration. In what follows the (<) configuration will be referred to as a *double probe link* and (–) as a *single probe link.*

Within each of the problem groups RT increases with the number of links. If a sequential processing of links is assumed then the slopes of the true curves directly reflect the average time taken to verify that a link is true. Note that the three true curves plotted in Fig. 3 have approximately the same slopes, which together with the equalities remarked above is consistent with a sequential processing account. A way to investigate order in sequential processing is to examine false RT's for each problem type in a group as a function of the position of the false link assuming that subjects responded "false" as soon as they discovered a false link. Figure 4 illustrates graphically this order of processing analysis. Figure 4a shows that for group A problems RT increases as the false link is moved from the first to the third position with a slope about the same as the true slopes in Fig. 3: this indicates a strict left/right processing order. Figure 4b shows that for P(–<) the tail link (–) is almost always verified before the left branch (<), while within the branch there is no strong up/down processing order. This is interpreted as consistent with the proposal that the double probe link is verified in one step (i.e., not as separate simple links) which implies that there should be no up/down processing order as such. Figure 4c presents a more complicated story for group C. While link processing for this group tends to be up/down on the right branch (>) and branch (>) before tail (–) (i.e., left/right) in P(>–), this order cannot be strict since the RT slopes as the false link position moves are noticeably less than the true slopes in Fig. 3. A probabilistic order of processing is appropriate for group C problems.

The verification of a link is in some respects similar to memory scanning tasks (Sternberg, 1969a) that require subjects to decide whether a probe symbol is contained in a memorized set of symbols. For an ASP transition table the number of elements in a list labelled by a cue word is referred to as the *cue set size*; Fig. 5 plots true and

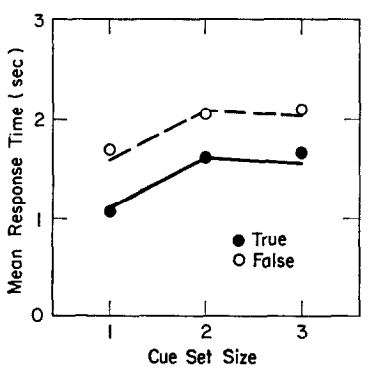


FIG. 5. Mean correct RT's for problem type P(–) plotted by cue set size and true/false.

false RT's for P(–) by cue set size to illustrate set size effects analogous to those found in memory scanning tasks. The true and false curves are separated by a constant, suggesting a simple additive effect on RT of the process differences between true and false link verifications.

Errors were infrequent under the speed/accuracy instructions given the subjects; the error rate over all conditions and subjects was 5.2 percent. Group error rates broken down by problem type and position of the false link are presented in Table I. While the authors recognize the possibility of important theoretical relations between response times and error rates as for example suggested by Pachella (1974) among others, a rigorous analysis relating the two was not performed for the data presented here. This omission is partly justified by the empirical observation that while mean RT's showed a consistent pattern across subjects, error rates did not. Also, from purely statistical considerations when data is so finely classified that some classifications have twenty or fewer observations, error rates may not be sufficiently reliable for the analysis of data from an individual subject whereas RT's may still be meaningful in providing insight into psychological processes.

Theoretical Analysis

Suppose that for a particular ASP item we have been given a description of the sequence of psychological operations used to solve it. The stage modeling technique to be used here assigns to each operations or stage, S, of the processing a tuple of parameters,

$$\langle \mu(S), \sigma^2(S) \rangle,$$

corresponding to the theoretical mean and variance of processing time associated with that stage. In cases more general than that considered here this tuple may become a family of tuples corresponding to various states of the cognitive system that could exist when the stage operates (i.e., stages are specified conditionally) or tuples may contain additional parameters such as higher RT moments or the probability of a processing failure in that stage. If stages S_1, S_2,..., S_m are applied in sequence to process the item then the RT mean and variance for the item are simply,

$$\mu(\mathrm{RT}) = \sum_{j=1}^{m} \mu(S_j) \qquad \text{and} \qquad \sigma^2(\mathrm{RT}) = \sum_{j=1}^{m} \sigma^2(S_j).$$

The additivity of variance follows from the assumption that stage processing times are stochastically independent. Now suppose that there are two sequences of stages that

could be applied to the item, S_{11}, S_{12},..., S_{1m_1} and S_{21}, S_{22},..., S_{2m_2}, and that these two sequences are observed with probability p and $(1 - p)$, respectively. Let,

$$\mu_i = \sum_{j=1}^{m_i} \mu(S_{ij}) \quad \text{and} \quad \sigma_i^2 = \sum_{j=1}^{m_i} \sigma^2(S_{ij}), \quad i = 1, 2.$$

Then the mean and variance of the overall RT are

$$\mu(\text{RT}) = p\mu_1 + (1 - p)\,\mu_2\,,$$
$$\sigma^2(\text{RT}) = p\sigma_1^2 + (1 - p)\,\sigma_2^2 + p(1 - p)(\mu_1 - \mu_2)^2.$$

Without going into further detail, similar expressions can be derived whenever RT is assumed to arise from the probabilistic mixture of sequences of stages.

Proceeding on the basis of the observations made in the Results section above, a stage model was constructed using a small number of inclusive stages that are identifiable (i.e., in the sense of unique parameter estimates) and that have direct theoretical interpretation. These stages are

Stage	Stage description
V_n	verification of a single probe link with cue set size of n
W_n	verification of a double probe link with cue set size of n
K	orientation, attention, perception and miscellaneous set-up and bookeeping processes
D	decision and response processes that differ between "true" and "false" responses.

Processes involved in the verification of single and double probe links have been summed together in the V_n and W_n parameters, respectively. Due to the problem of identifying parameters it is not possible to make definitive interpretations of the stages K and D. The K stage includes all those operations which are in common across problem items, such as attending to the CRT display or executing the motor components of a keypress response; in addition, K may be regarded as incorporating incidental processes required for the logical completeness of the model such as recording the input and output of stage operations. Any processing differences between true and false responses, including handedness, are incorporated in the D stage. For the experimental data false responses are slower than comparable true responses; the D parameters reflect this aspect of the data.

Derivation of Theoretical Expressions

The derivation of expressions for theoretical RT means and variances will be illustrated by examples since there is insufficient space for an exhaustive treatment. In the following, let n_i be the cue set size associated with the symbol x_i.

EXAMPLE 1. $x_i \rightarrowtail x_j$ (T). To solve this simplest problem the subject need only verify one link; hence, exactly the stages V_{n_i} and K occur. Then,

$$\mu(\mathrm{RT}) = \mu(V_{n_i}) + \mu(K),$$
$$\sigma^2(\mathrm{RT}) = \sigma^2(V_{n_i}) + \sigma^2(K).$$

EXAMPLE 2. $x_i \rightarrowtail x_j \rightarrowtail x_k \rightarrowtail x_l$ (TFT). Assuming that P(———) has a strict left/right processing order, the subject first verifies that $x_i \rightarrowtail x_j$ is true and then finds that $x_j \rightarrowtail x_k$ is false; the subject responds "false" as soon as she finds this link so that only stages V_{n_i}, V_{n_j}, K, and D occur. Then,

$$\mu(\mathrm{RT}) = \mu(V_{n_i}) + \mu(V_{n_j}) + \mu(K) + \mu(D),$$
$$\sigma^2(\mathrm{RT}) = \sigma^2(V_{n_i}) + \sigma^2(V_{n_j}) + \sigma^2(K) + \sigma^2(D).$$

EXAMPLE 3. $x_i \rightarrowtail x_j$ ($x_j \rightarrow x_k$, $x_j \rightarrow x_l$) (TFT).

It is assumed that the double probe link ($<$) is verified in a single operation, W_{n_j}, and that the tail (–) is checked before the branch ($<$), so that the stages are V_{n_i}, W_{n_j}, K, and D. Then,

$$\mu(\mathrm{RT}) = \mu(V_{n_i}) + \mu(W_{n_j}) + \mu(K) + \mu(D),$$
$$\sigma^2(\mathrm{RT}) = \sigma^2(V_{n_i}) + \sigma^2(W_{n_j}) + \sigma^2(K) + \sigma^2(D).$$

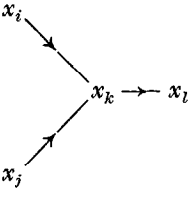
EXAMPLE 4. ($x_i \rightarrow x_k$, $x_j \rightarrow x_k$) (FT).

A probabilistic order of processing was suggested for types P($>$) and P($>$–). This order will be defined by two probability parameters. Let q be the probability that within a right branch ($>$) the upper link is checked before the lower link, and let r be the probability that for P($>$–) the branch ($>$) is checked before the tail (–). In Example 4 the parameter r is not involved. With probability q the stages are V_{n_i}, K,

and D, and with probability $(1 - q)$ the stages are V_{n_i}, V_{n_j}, K, and D, with the result that

$$\mu(\mathrm{RT}) = \mu(V_{n_i}) + (1 - q)\,\mu(V_{n_j}) + \mu(K) + \mu(D),$$
$$\sigma^2(\mathrm{RT}) = \sigma^2(V_{n_i}) + (1 - q)\,\sigma^2(V_{n_j}) + \sigma^2(K) + \sigma^2(D) + q(1 - q)[\mu(V_{n_j})]^2.$$

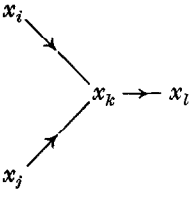
The expression for $\sigma^2(\mathrm{RT})$ is that for the probability mixture of two sequences of stages.

EXAMPLE 5.  (TFT).

In Example 5, with probability rq the stages are V_{n_i}, V_{n_j}, K, and D; with probability $r(1 - q)$, V_{n_j}, K, and D; with probability $(1 - r)q$, V_{n_k}, V_{n_i}, V_{n_j}, K, and D; and with probability $(1 - r)(1 - q)$, V_{n_k}, V_{n_j}, K, and D. Hence,

$$\mu(\mathrm{RT}) = q\mu(V_{n_i}) + \mu(V_{n_j}) + (1 - r)\,\mu(V_{n_k}) + \mu(K) + \mu(D),$$
$$\sigma^2(\mathrm{RT}) = q\sigma^2(V_{n_i}) + \sigma^2(V_{n_j}) + (1 - r)\,\sigma^2(V_{n_k}) + \sigma^2(K) + \sigma^2(D) + q(1 - q)[\mu(V_{n_i})]^2 + r(1 - r)[\mu(V_{n_k})]^2.$$

The expression for $\sigma^2(\mathrm{RT})$ is an algebraic simplification of a general expression.

EXAMPLE 6.  (TTT).

Since all links are true, the same stages must occur whatever the order of processing. Consequently,

$$\mu(\mathrm{RT}) = \mu(V_{n_i}) + \mu(V_{n_j}) + \mu(V_{n_k}) + \mu(K),$$
$$\sigma^2(\mathrm{RT}) = \sigma^2(V_{n_i}) + \sigma^2(V_{n_j}) + \sigma^2(V_{n_k}) + \sigma^2(K).$$

These examples should convey the gist of the statistical model. Note that for every

ASP item the theoretical RT mean and variance can be expressed in the following canonical forms:

$$\mu(\mathrm{RT}) = a_1\mu(V_1) + a_2\mu(V_2) + a_3\mu(V_3) + a_4\mu(W_2) + a_5\mu(W_3) + a_6\mu(K) + a_7\mu(D)$$
$$\sigma^2(\mathrm{RT}) = a_1\sigma^2(V_1) + a_2\sigma^2(V_2) + a_3\sigma^2(V_3) + a_4\sigma^2(W_2) + a_5\sigma^2(W_3) + a_6\sigma^2(K) + a_7\sigma^2(D) + b^2.$$

where the a_i's ($i = 1,..., 7$) can be interpreted as the average number of times the corresponding stage occurs, and b^2 is the "mixture variance" (i.e., the variance added by mixing processing strategies where strategies may require differing amounts of time). Writing the row vectors,

$$\mathbf{e} = \langle \mu(V_1), \mu(V_2), \mu(V_3), \mu(W_2), \mu(W_3), \mu(K), \mu(D) \rangle,$$
$$\mathbf{v} = \langle \sigma^2(V_1), \sigma^2(V_2), \sigma^2(V_3), \sigma^2(W_2), \sigma^2(W_3), \sigma(K), \sigma^2(D) \rangle,$$
$$\mathbf{a} = \langle a_1, a_2, a_3, a_4, a_5, a_6, a_7 \rangle,$$

the canonical forms become,

$$\mu(\mathrm{RT}) = \mathbf{a}\mathbf{e}^{\mathrm{T}} \quad \text{and} \quad \sigma^2(\mathrm{RT}) = \mathbf{a}\mathbf{v}^{\mathrm{T}} + b^2$$

where $\mathbf{e}^{\mathrm{T}}$ is the transpose of $\mathbf{e}$, and $\mathbf{v}^{\mathrm{T}}$ is the transpose of $\mathbf{v}$. For all true items and for false items in groups A and B, each a_i is an integer and $b^2 = 0$; for false items in group C the a_i's may be functions of q and r, and $b^2 > 0$ is a function of $\mathbf{e}$, q and r. Note that ASP items can be classified according to their coefficient vectors, $\mathbf{a}$, and mixing variances, b^2; under the model this classification is a full specification of the items. For the items used in the experiment 46 such classification categories occurred.

Statistical Evaluation

A discussion of parameter estimation and statistical techniques is presented in the appendix. Best estimates of parameters were obtained for each subject by numerical methods using a quadratic loss function, and the fit of the model to the RT data was primarily evaluated by constructing simultaneous confidence regions containing all the RT means and variances predicted by the model. Parameter estimates are given in Tables II and III; statistics are listed in Tables IV and V.

For the mean RT data the statistics in Table IVA show that while the model does account for a substantial percentage of the between and total variances (PBV and PTV columns), the maximum modulus t test applied to the group suggests that the model is probably not a complete account of the data for every subject in the experiment (g^* for the group is .004 which is the probability of observing a t^* value of 4.44 or greater). In Table IVA two of six subjects have $g^* > .10$ indicating a good fit of the model for these individual data, and in Table IVB $g^* > .10$ for three subjects. The third column of Table IV gives the number of points lying outside a .90 simultaneous confidence interval; any such point implies that $g^* < .10$.

TABLE II

Parameter Values, Averaged Estimates[a]

	Stage	μ (msec)	σ (msec)
Single probe (–) verification:			
Cue set size 1	$V1$	676	241
Cue set size 2	$V2$	1169	446
Cue set size 3	$V3$	1146	563
Double probe (<) verification:			
Cue set size 2	$W2$	924	365
Cue set size 3	$W3$	1206	412
Set-up processes	K	446	210
True/false difference	D	466	200

[a] Probability of up before down on (>), $q = 0.85$; probability of (>) before (–) on (>–), $r = 0.89$.

TABLE III

Parameter Values, Individual Subjects

	Subject 1		Subject 2		Subject 3	
	μ	σ	μ	σ	μ	σ
$V1$	878	566	449	3	571	0
$V2$	1545	676	867	293	1001	450
$V3$	1587	720	913	593	775	446
$W2$	968	575	789	399	748	398
$W3$	1144	442	986	421	768	284
K	475	1	545	1	394	0
D	729	1	346	489	452	0
q	0.79		1.00		0.78	
r	0.88		0.48		1.00	
	Subject 4		Subject 5		Subject 6	
$V1$	956	0	594	259	607	532
$V2$	1375	671	1430	635	798	230
$V3$	1348	729	1271	646	980	636
$W2$	1156	689	1045	31	838	78
$W3$	1582	720	1455	485	1301	88
K	292	1	462	307	506	414
D	459	1	302	2	506	1
q	0.61		0.92		1.00	
r	1.00		0.99		1.00	

TABLE IV

Statistical Analysis for RT Means

Subject	t^*	g^*	Points outside 90% region	*PBV*	*PTV*
	A. Model classification (46 points)				
1	4.44	0.000	2	79.7	42.7
2	2.78	0.182	0*	76.7	35.6
3	3.32	0.029	2	78.6	40.7
4	4.44	0.000	1	79.3	36.5
5	3.86	0.007	2	57.7	28.1
6	2.84	0.157	0*	68.7	27.8
Group	4.44	0.004	3	73.5	35.2
	B. Type *X* false link classification (23 points)				
1	3.28	0.014	1	85.6	41.5
2	1.74	0.697	0*	84.3	32.1
3	2.42	0.198	0*	89.5	37.3
4	3.01	0.036	1	83.7	35.8
5	3.44	0.008	1	64.5	26.6
6	2.67	0.101	0*	71.0	24.5
Group	3.44	0.049	2	79.8	33.0

TABLE V

Statistical Analysis for RT Variances[a]

Subject	t^*	g^*	Proportion model vs "same"
1	0.83	0.999	0.478
2	0.63	0.999	0.500
3	1.00	0.999	0.348
4	1.07	0.999	0.522
5	0.85	0.999	0.478
6	0.56	0.999	0.391

[a] Model classification, 46 points.

Table V presents statistics for RT variances. Due to large sampling variability of variance estimates, the g^* statistics are not very informative since many models would be acceptable within wide limits of variability. The third column of Table V compares the model to the hypothesis that all RT variances are the same, in terms of the proportion of points for which the model makes a more accurate prediction. Evaluated using this statistic the model does no better than the "same" hypothesis although both are acceptable given the variance of estimators. Since the averaged RT variance parameters presented in Table II appear to be orderly they will be discussed, although no strong conclusions should be drawn.

Discussion

Stage modeling has been conceived of in terms of a formal processing language description of memory operations: stages are analogous to procedures or subroutines, perhaps probabilistic in their execution, organized by call sequences into memory processes. Within such a stage modeling framework various levels of detailed description are possible. For example, one might consider macro stages such as "perception," "memory," and "response," or comparatively micro stages such as "input the symbol in position p of the stimulus array" or "compare the code for symbol X with the code for symbol Y." No particular level of detail can be regarded as preferred: theoretical descriptions in stage terms must be evaluated with respect to the relevant data. However, the stage modeling framework does in principle relate all levels of description in terms of the nesting of procedures in call sequences, thus providing the possibility of consistently treating the results of simple and relatively more complex laboratory tasks with the same overall processing model (Atkinson & Wescourt, 1975).

A stage model can be most productively regarded as a rational basis for the construction of statistical models. Each statistical model stemming from a stage model can be evaluated with respect to the data, successes and failures yielding new information about the data possibly not apparent on inspection or available from other analyses. In general it is not necessary that every statistical model derived from a particular processing language description be "successful," but only that some are, and that these provide a useful characterization of the data. Of course, if a stage model were taken as a literal model of a specific real time process, say specific interactions among brain centers and layers of brain tissue, it would be important to verify all the statistical models derived from the stages theory. However, for the analysis of cognitive performance the authors regard stage models as nonliteral information processing descriptions from which statistical analyses are derived that provoke a deeper and more adequate characterization of patterns present in the data.

The statistical models for RT means and variances developed above may be regarded as an intermediate level of stages analysis appropriate to the level of observable data:

It does not explicitly describe either the component processes of individual link verifications or the overall control structure in which the problem solving algorithms are embedded. Since these additional levels of analysis are of theoretical interest, the discussion will turn to bridging these conceptual gaps. The following stages analysis of single and double probe link verification is given:

L_0 determine whether single or double probe (s or d)

L_1 input cue

L_2 access memory list associated with cue

L_3 input probe 1; if d, then input probe 2

L_4 reset match register 1; if d, then reset register 2

L_5 unpack an element from the memory list

L_6 match the element against probe 1 and increment match register 1 by the value of the "goodness-of-match"; if d, then match against probe 2 and increment register 2

L_7 if the entire list has been unpacked then continue, else return to L_5

L_8 if d, then add match register 2 to register 1

L_9 if s, then if the value of match register 1 exceeds a criterion c_s then return true, else return false; if d, then if the value exceeds c_d return true, else return false

Note that the analysis is essentially an "exhaustive scan" model, where matching is not necessarily all-or-none, and where the representation of lists in memory and the coordinate retrieval or unpacking process may be more involved than reading from a list of symbols at a uniform rate. Representing a list as a cluster of symbols bound to a memory node by associative linkages and defining retrieval processes in terms of this representation would be one way of conceiving of an unpacking operation with more complex characteristics, although such "built-in" characteristics may have limited conceptual and theoretical interest.

The claim is that this model of link verification is consistent with the stage parameter estimates in Table II; for the sake of simplicity only the average values of parameter estimates are discussed. The parameter values in Table II may be qualitatively summarized as follows:

$$\begin{aligned}
&\mu(V_1) < \mu(V_2) = \mu(V_3),\\
&\mu(W_2) < \mu(W_3),\\
&\mu(W_2) < \mu(V_2),\\
&\mu(W_3) = \mu(V_3),\\
&\sigma^2(V_1) < \sigma^2(V_2) < \sigma^2(V_3),\\
&\sigma^2(W_2) < \sigma^2(W_3),\\
&\sigma^2(W_2) < \sigma^2(V_2),\\
&\sigma^2(W_3) < \sigma^2(V_3).
\end{aligned}$$

This summary can be regarded as a hypothesis that, within the sampling variability of the parameter estimates, is not disconfirmed. A problematic aspect of this summary is that $\mu(W_2) < \mu(V_2)$ by 245 msec, yet $\mu(W_3) = \mu(V_3)$. This result may be attributable in some way to the fact that for W_2 the number of probes is the same as the cue set size, but in the absence of additional controls no ad hoc explanations are offered.

If it is assumed that stages L_2, L_5, and L_9 account for the major part of link verification time, then a gross similarity would be predicted between single and double probe links. With suitably complex representations of lists the mean unpacking time for lists of lengths 2 and 3 may be comparable, yielding $\mu(V_2) = \mu(V_3)$; the speed of V_1 could be explained by the simplicity of the representation for a list with a single symbol requiring fewer unpacking manipulations.

Single and double probe link verifications differ in stages L_8 and L_9. If the matching process is probabilistic (e.g., due to variable imperfect coding of symbols) then the final match value in register 1 will be distributed differently for single and double probes (e.g., double probes will have greater mean and variance for both true and false links). This, together with the two criteria, c_s and c_d, might account for decision component differences in ways similar to signal detection models that relate RT to criteria placements in relation to signal and noise distributions (Thomas, 1971). The observed $\sigma^2(W_2) < \sigma^2(V_2)$ and $\sigma^2(W_3) < \sigma^2(V_3)$ are interpreted as due to such differential effects in stage L_9.

The value $\mu(D) = 466$ is greater than would be expected on the basis of handedness alone, suggesting genuine decision component differences; again this is interpreted as a stage L_9 effect. Since successive link verifications are required by some ASP items, in order to achieve an acceptable error rate (subjects were instructed to be accurate) it is necessary to make a more accurate decision for each intermediate verification than would be needed if only a single link were verified on each trial. Also, since over all items there are more true than false links, stage L_9 might be "tuned" for a true verification. The demand for increased accuracy together with a true verification expectancy could account for the observed value of $\mu(D)$. The apparent constancy of $\mu(D)$ over problem types, even those where only one link is verified, is consistent with the theoretical conception that the same link verification mechanisms are used for all problems without modification according to problem type. From these considerations it would be predicted that encouraging speed over accuracy, using only single link problems, reducing the variety of ASP items used within an experiment, or using multilink items with more than one false link would all have an effect in reducing the value of $\mu(D)$.

As an aside, it may be possible to use an empirical speed/accuracy tradeoff to further investigate the verification mechanisms found in the ASP task. A direct implication of the theory discussed above is that under speed instructions each link verification will be less accurate as processing is modified for speed or cut short, with the results that errors will tend to increase relatively more for items with many links compared to

those with few, and that error RT's for multilink true items will decrease relative to correct RT's while error RT's for multilink false items will increase. Other, quite different effects of speed instructions might be to induce subjects to implement faster problem solving algorithms, say with some sort of simultaneous verification of links, to "prime" access to certain algorithms and retrieval mechanisms in anticipation of the next problem, or to adopt sophisticated guessing strategies. The issues with regard to speed/accuracy effects in ASP problem solving are manifold and may perhaps be most productively approached by comparing results across experiments to determine what effects might be present.

In stage terms a *stable strategy* is a problem solving algorithm that is not modified with use. Empirically, stable strategies would be expected for practiced subjects who have in some sense developed optimal task techniques, with the required amount of practice depending on the particular task. The present experiment was designed to observe only asymptotic performance, making it in principle possible to specify a single set of algorithms or strategies governing the processing of ASP items. A theory as to how these strategies are set up with practice is not developed here; however, the authors do conceptualize an interactive feedback system where the state space of the system consists of algorithms and the effects of control inputs are to rebuild algorithms. It is proposed that for tasks where alternative processing strategies are a genuine theoretical possibility, it may be more appropriate to analyze data from trials early in the experiment in terms of a mixture of strategies rather than a single stable strategy. For the sake of completeness of theoretical conception it is assumed that the strategies for the various problem types are called by a controlling stage that on each trial identifies the problem type on the basis of its mapping diagram configuration and calls the corresponding problem solving algorithm.

Additional empirical work is required to evaluate these conceptual analyses of control and component processes. For example, one line of experimentation would be to examine more thoroughly ASP verification problems, with manipulations of the transition table and problem types. Another line would be to examine ASP problems more complex than verification, with the idea that such tasks could reveal more about the construction of strategies, that is about how component processes are used to build problem solving algorithms. Alternatively, the verification of isolated single and double probe links could be examined in greater experimental detail. All these levels of experimental investigation are well integrated within the stage modeling framework, which is one of the main theoretical motivations for using such a framework as a basis for data analysis.

From a theoretical standpoint a close relationship exists between link verification and some memory scanning tasks. In both cases a probe item must in some sense be compared against a list of symbols in memory to determine if the probe is a member of the list. A point of interest is whether memory scanning mechanisms that have been investigated in the laboratory can be identified as components of relatively more

complex tasks such as solving ASP verification problems. The model constructed for the ASP problems investigated in this paper can be regarded as an attempt to tackle this issue. About the simplest relation that could obtain between memory scanning and ASP problem solving would be that the scanning mechanisms engaged by strategies to yield intermediate results have the same characteristics as those observed with simple memory scanning tasks. Yet this need not be so. It is conceivable that as strategies for the more complex storage, retrieval and decision making required by ASP problems are constructed in the memory system (Atkinson & Wescourt, 1975), new demands for rapid access to a larger volume of stored information, for the recording of intermediate results which direct further processing, and for controlling error rates when intermediate results are combined or cascade in a final decision demand scanning mechanisms having different characteristics. The data from the present experiment are not in themselves conclusive, but the parameter values of Table II as discussed above suggest that the inferred scanning (link verification) mechanisms and decision processes yield values of RT parameters that differ from those typically found in the memory scanning literature. There is the unexpected result that verifying a double probe link is as fast as verifying a single probe link; the fact that for single probe links verification times for cue set sizes 2 and 3 do not differ from each other but are dramatically different from the verification time for cue set size 1; and the unusually large constant difference between true and false RT's. Each of these effects is of course subject to further investigation and taken one at a time are not without some parallel in the memory literature, but the authors believe that they provoke an examination of the issue of how memory scanning mechanisms relate to the larger human memory system. It is fair to say that proportionally more effort has been devoted to unraveling the effects of experimental manipulations on basic memory scanning tasks and constructing sophisticated and interesting models for these data (e.g. Theios, 1972; Anderson, 1973; Shevell & Atkinson, 1974) than has been devoted to examining the possible roles of memory scanning mechanisms in human memory systems that are sufficient to support more involved cognitive processing.

The stage model developed for the experiment described here characterized each stage by two parameters, the mean and variance of processing time; as remarked above, this type of model can be generalized to include more parameters such as the probability of an error in that stage or higher moments of the processing time distribution. Without changing the nature of the modeling technique, stage parameters could be expressed conditionally on the state of processing, as for example on the input to the stage from previously operating stages. Even with these generalizations, parameter estimation and statistical procedures can be derived in a mathematically simple way. Granted that it is one opinion, the authors feel that statistical methods such as those described in this paper that are based on a formal but flexible model of psychological processing should in many cases be both practical and more incisive than the standard linear statistical analyses often found in the memory and problem solving literature.

Comparison with Hayes' spy problems

Hayes (1965, 1966) has reported studies using a problem solving task similar to that of the ASP problems defined here. Subjects in Hayes' experiments learned a list of "spy" names together with rules about which spies could talk to each other; the list of these "talking connexions" may be regarded as a transition table. In the basic experiment, subjects were given two spy names and required to find a chain of spy-to-spy communications conveying a message from the one spy to the other. Subjects were instructed to "think aloud" and their protocols were analyzed with respect to the overall time taken to solve a problem, the rate at which links in the communication chain were generated, and diversions into "blind alley" side chains (i.e., passing the message to a spy who did not have the connections to get it to the goal spy). Subjects were able to solve spy problems in a matter of a few minutes, occasionally entering side chains and usually achieving a solution chain longer than the minimal required chain; the reader is referred to the original papers for Hayes' analysis of his results. In terms of the type of theory proposed here for ASP problems, the solution of spy problems would be described by algorithms constructed using a small set of basic psychological operations and following specific search-and-test methods of chain construction. Insofar as the model stated definite algorithms it would have the potential to account for protocols; as stage models the algorithms would also make quantitative predictions about the pattern of observed RT's and error rates. Of course the particular theory of ASP problem solving outlined in this paper is not sufficient in itself to account for Hayes' results such as the end-acceleration phenomenon: In addition, explicit algorithms would have to be constructed and demonstrated by computer simulation or by inferential data analysis to produce the observed pattern of results.

The Stage Modeling Technique

It is worthwhile to emphasize the positive aspects of stage modeling as a technique for the analysis of RT tasks. Interesting arguments related to those presented here have been given by Sternberg (1969b) with respect to the so-called additive factors method. First, as has been noted, considering psychological processes as procedures or subroutines in the sense of a formal computer language provides an easily conceived unifying framework for theoretical analysis and a rationale for investigating memory mechanisms as they occur both in simple and complex laboratory tasks. Second, from a statistical standpoint, regression models for RT moments can be derived from a stages theory in a relatively simple manner, basically by counting the occurrences of stages. The parameters in the regression model have direct psychological interpretation in terms of real processing time, and the parameters can be estimated by common analytic or numerical methods irrespective of the number of classification categories or the number of observations in each category. With regard to predictive power, stage models can provide accounts for RT moments of all orders and together with

notions of processing variability defined at specific stages can at the same time provide an account of errors. Even though the technique is mathematically simple, the underlying process representation is that of a quite general sequence of random variables (or random vectors) corresponding to the definition of a discrete stochastic process (viz., a family of random variables with a countable index set) with very few restrictions (e.g., most of the random variables can be assumed to be finite valued). This suggests that many models of memory processes will be at least formally "nearly" equivalent to some stage model as defined here. The nature of this equivalence can be formalized in terms of the partitioning of the event space of the experiment (i.e., the set of all possible data points) induced by the inverse mapping of the goodness of fit measure regarded as a random variable.

Simple and Complex Tasks

The algebra step problems introduced in this paper are, like other artificial memory and problem solving tasks, not advocated for their intrinsic interest but rather as one experimental paradigm for testing our understanding of human memory systems. Fast accurate problem solving has on the one hand clear theoretical relations to conceptions of basic memory mechanisms and the manner in which these mechanisms come to play in a larger memory system, and on the other hand it is a bridge to the chronometic analysis of more traditional problem solving tasks. While the investigation of simple tasks is indispensable it is surely necessary to develop theoretical constructions for more complex tasks with equal vigor: the chronometric analysis of tasks at the level of ASP problems is intended as one step in this direction. In philosophical perspective there is no assurance that even a detailed understanding of the models required to account for isolated simple memory tasks will automatically lead to an adequate conception of human memory systems that are capable of supporting such routine cognitive functions as the retrieval of propositional information (Anderson & Bower, 1973) or grade school arithmetic problems (Suppes, Loftus & Jerman, 1969). The data and analysis presented in this paper suggest that analysis of RT's on the order of 5 sec is feasible without undue loss of precision either in the conceptual model or the statistical treatment. Across experiments it should be possible to identify the characteristics of memory mechanisms as they occur in memory systems where processes involving alternative strategies, intermediate processing results and decisions about subsequent processing, and rapid access to large amounts of stored information are operating. Such a program of research has the potential to develop the basis for more exacting analyses of problem solving tasks in terms of an explicit theory of human memory, to elucidate the role of control and decision processes, and to qualify our understanding of memory mechanisms discovered through research on simple tasks.

Statistical Appendix

The coefficient vectors $\mathbf{a} = \langle a_1, \ldots, a_7 \rangle$ define a classification of observations into distinct categories under the model; 46 such *classification categories* were observed in the experiment (i.e., there were 46 distinct $\mathbf{a}$ vectors). The notation below will be used in what follows: the index "i" refers to the ith subject and "j" to the jth classification category.

n	number of classification categories under the model,
s	number of subjects,
N_{ij}	number of observations,
M_{ij}	RT sample mean,
M_{i+}	RT grand sample mean,
S_{ij}^2	RT sample variance,
T_{ij}^2	sample variance of S_{ij}^2 (see Kendall & Stuart, 1969).

Parameter Estimation

The approach taken to parameter estimation was to choose a *loss function* conceived of as a function of the parameters given the data, and to find parameter values that minimized this function. Since function minima were found using a numerical grid search method, computationally efficient quadratic (least squares) loss functions were chosen. Parameters were estimated for each subject individually.

The actual estimation proceded in two steps. First, values of $\mathbf{e}_i$, q_i, and r_i were determined using the loss function

$$\mathrm{LS}_1(\mathbf{e}_i, q_i, r_i) = \sum_{j=1}^{n} N_{ij}[(M_{ij} - \mathbf{a}_{ij}\mathbf{e}_i^{\mathrm{T}})^2/S_{ij}^2].$$

Second, the parameter values, $\hat{\mathbf{e}}_i$, $\hat{q}_i$, $\hat{r}_i$, were treated as constant and $\mathbf{v}_i$ estimated with the loss function

$$\mathrm{LS}_2(\mathbf{v}_i \mid \hat{\mathbf{e}}_i, \hat{q}_i, \hat{r}_i) = \sum_{j=1}^{n} \{[S_{ij}^2 - (\mathbf{a}_{ij}\mathbf{v}_i^{\mathrm{T}} + b_{ij}^2)]^2/T_{ij}^2\}.$$

An alternative procedure would have been to estimate simultaneously all parameters using a combined loss function of the form,

$$\mathrm{LS} = w\,\mathrm{LS}_1 + (1 - w)\,\mathrm{LS}_2, \qquad 0 < w < 1.$$

However, it was observed that the RT means showed a clearer pattern than the RT variances, so that estimates of the mean RT parameters "uncontaminated" by possible failures of the model for RT variances were considered appropriate.

Parameter estimates for individual subjects are listed in Table III. The numerical method used to estimate variance parameters excluded negative variances with one result that some parameters were estimated to be near zero (the loss function, LS_2, would have been reduced had negative values been accepted for these parameters). An inherent problem in the analysis of RT variances is that for classification categories with small sample sizes, the variability of the sample variance, S_{ij}^2, is large relative to that for the sample mean, M_{ij} : Consequently, parameter estimates will also have large variability. Note that variance parameter estimates averaged across subjects are more readily interpretable as variability is reduced through averaging.

Goodness of Fit Measures

Consider the statistic defined for the ith subject and jth category by

$$t_{ij} = N^{1/2}(M_{ij} - \mathbf{a}_{ij}\hat{\mathbf{e}}_i^T)/S_{ij},$$

which for suitable models may be assumed to be approximately distributed as Student's t under the hypothesis that the theoretical mean, $\mathbf{a}_{ij}\hat{\mathbf{e}}_i^T$, is the true mean of the ijth RT distribution. One method of evaluating the fit of the model to mean RT's is to construct the smallest possible uniform simultaneous confidence region containing all the t_{ij}'s and to note the probability of the complement of the region. This probability is the minimum value of α (the probability of a type I error) for which the hypothesis that the model is true can be rejected; small values indicate that the model is probably not a full account of the mean RT data. If the distribution of t_{ij} is approximated by $N(0, 1)$ instead of by Student's t, a conservative bias is introduced in the sense that the value of α is necessarily reduced. Since the normal approximation simplifies the calculation of a simultaneous confidence region this assumption is adopted.

For the ith subject define

$$t_i^* = \max_{1 \leqslant j \leqslant n} | t_{ij} |.$$

If the t_{ij}'s were independent, then for any positive number, c,

$$\begin{aligned} \Pr\{t_i^* > c\} &= 1 - \Pr\{t_i^* < c\} \\ &= 1 - \prod_{j=1}^{n} \Pr\{-c < t_{ij} < c\} \\ &= 1 - [\Pr\{-c < z < c\}]^n, \end{aligned}$$

where $z \sim N(0, 1)$. This is the probability that for a fixed i all the t_{ij}'s are contained within a uniform, symmetric confidence band of width $2c$. But for each i the t_{ij}'s are correlated through the estimation procedure, and with enough parameters it may be possible to obtain all the $t_{ij} = 0$, rendering the preceding probability statements meaningless. Accordingly, some conservative adjustment should be made taking into

account at least the number of free parameters, p. The choice for the present analysis was to take

$$g_i^* = 1 - [\Pr\{-c < z < c\}]^{(n-p)}$$

in place of $\Pr\{t_i^* > c\}$ above. If $t_i^* = c$ is observed, then g_i^* is a statistical measure of the fit of the model for the ith subject. Similarly, for a sample of s subjects define

$$t^* = \max_{1 \leqslant i \leqslant s} | t_i^* | = \max_{\substack{1 \leqslant i \leqslant s \\ 1 \leqslant j \leqslant n}} | t_{ij} |;$$

then

$$g^* = 1 - [\Pr\{-c < z < c\}]^{s(n-p)}$$

is a goodness of fit measure for the sample as a whole. This procedure is a type of multiple modulus test (Miller, 1966) referred to here as a "maximum modulus t test" with $(n - p)$ or $s(n - p)$ "degrees of freedom," taking some licence with terminology.

A related procedure can be followed in evaluating theoretical versus observed RT variances. The statistics defined by

$$t'_{ij} = \frac{S_{ij}^2 - (\mathbf{a}_{ij}\hat{\mathbf{v}}_i^{\mathrm{T}} + b_{ij}^2)}{T_{ij}}$$

can be treated in the same manner as the t_{ij}'s above, although t'_{ij} cannot be regarded as having Student's t distribution and g^* in this case ought to be taken as a transformation of the t'_{ij}'s reflecting goodness of fit rather than as an approximation to a true probability.

To obtain a firmer statement about goodness of fit a second measure was sought. Although the model under consideration is not linear, the total sum of squares can be partitioned in such a way as to yield statistics reflecting the goodness of fit of the model to RT means in a way similar to the percentage of between variance accounted for and the sample correlation coefficient in linear regression. Define for any set of theoretical means, $\{f_{ij}\}$, for the ith subject:

$$MV_i = \mathrm{SS(between)} - \sum_{j=1}^{n} N_{ij}(f_{ij} - M_{ij})^2 - 2 \left| \sum_{j=1}^{n} N_{ij}(f_{ij} - M_{ij})(f_{ij} - M_{i+}) \right|,$$

$$MV_i' = \max\{MV_i, 0\},$$

$$PBV_i = 100 \frac{MV_i'}{\mathrm{SS(between)}},$$

$$PTV_i = 100 \frac{MV_i'}{\mathrm{SS(total)}}.$$

If the f_{ij}'s were determined under a linear regression model using least squares estimation, then

$$MV_i' = MV_i = \text{SS (linear regression)}.$$

The results of the maximum modulus t, PBV and PTV analyses for RT means are presented in Table IV. Table IVA gives these statistics for the classification categories determined by the **a** vectors of the model; Table IVB represents the same analysis applied to the classification of Table I (problem type X position of false link). From Table IVA it is clear that the model accounts for a fair proportion of the variance (average PBV is 73.5 and average PTV is 35.2), yet only two subjects have $g^* > .10$, which is a "reasonable" criterion for a good fit. Additional information about the maximum modulus t test is given by the number of points falling outside the .90 confidence region; $g^* > .10$ if and only if this number is zero. It should be noted that points which lie outside the confidence region are not necessarily those which the model fails to account for since when parameters are estimated simultaneously for all points an "exceptional" point can adversely influence the prediction for other "normal" points. For the group of six subjects the maximum modulus t test indicates that the model is true can be rejected for $\alpha = .004$. It should be noted that one bad data point for a single subject can be sufficient to reject the model for the group using the maximum modulus t test; the proportion of subjects for which the model is not rejected is perhaps a more appropriate group statistic. In view of the all too common practice in the literature of presenting statistics for averaged group data, it is difficult to make a firm statement on this point based on the results of other comparable analyses.

The analysis presented in Table IVB indicates a slightly better fit although it is derived from a less strict interpretation of the model. Some improvement is expected since more extensive averaging may cancel out effects not accounted for by the model and estimated error variance is increased slightly as observations with different means are pooled. However, this second classification does correspond to an intuitively natural division of the data.

Table V presents an evaluation of the model's success in accounting for RT variances. As remarked above the variance of S_{ij}^2 is large for small sample sizes: for the experimental data this renders the maximum modulus t test uninteresting because for individual subjects the T_{ij}^2's are too large to reject any set of ballpark estimates for the variances. Variance predictions under the model were compared to the hypothesis that all the S_{ij}^2's are the same, using the proportion of points better accounted for by the model (absolute differences between predicted and observed were compared). Referring to Table V, the model succeeds about as well as the "same" hypothesis for four subjects and does worse for the remaining two subjects' data. This is not strong support for the model applied to RT variances but may be interpreted to mean that, compared to the "same" hypothesis, attempting to infer stage variances did not cost

much in the way of goodness of fit, while at the same time the model's predictions cannot be rejected given the variability of the S_{ij}^2 estimates.

REFERENCES

ANDERSON, J. A. A theory for the recognition of items from short memorised lists. *Psychological Review*, 1973, **80**, 417–438.

ANDERSON, J. R., & BOWER, G. H. *Human associative memory*. Washington, D.C.: Winston & Sons, 1973.

ATKINSON, R. C., & WESCOURT, K. T. Some remarks on a theory of memory. In *Attention and performance. V*. New York: Academic Press, 1975.

CHASE, W. G., & CLARK, H. H. Mental operations in the comparison of sentences and pictures. In L. W. Gregg (Ed.), *Cognition in learning and memory*. New York: Wiley, 1972.

DE GROOT, A. *Thought and choice in chess*. New York: Basic Books, 1965.

HAYES, J. R. Problem topology and the solution process. *Journal of Verbal Learning and Verbal Behavior*, 1965, **4**, 371–379.

HAYES, J. R. Memory, goals, and problem solving. In B. Kleinmuntz (Ed.), *Problem solving: Research, method, and theory*. New York: Wiley, 1966.

KENDALL, M. G., & STUART, A. *The advanced theory of statistics*, Vol. 1. New York: Hafner, 1969.

MILLER, R. J. *Simultaneous statistical inference*. New York: McGraw–Hill, 1966.

NEWELL, A., & SIMON, H. A. *Human problem solving*. Englewood Cliffs, N.J.: Prentice–Hall, 1972.

PACHELLA, R. G. The interpretation of reaction time in information processing research. In B. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition*. New York: Lawrence Erlbaum Associates, 1974.

Shevell, S. K., & Atkinson, R. C. A theoretical comparison of list scanning models. *Journal of Mathematical Psychology*, 1974, **11**, 79–106.

STERNBERG, S. Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*, 1969, **4**, 421–457. (a)

STERNBERG, S. The discovery of processing stages: Extensions of Donder's method. *Acta Psychologica*, 1969, **30**, 276–315. (b)

SUPPES, P., LOFTUS, E. F., & JERMAN, M. Problem solving on a computer based teletype. *Educational Studies in Mathematics*, 1969, **2**, 1–15.

THEIOS, J. Reaction time measurements in the study of memory processes: Theory and data. Report 72-2, Wisconsin Mathematical Psychology Program. Madison: University of Wisconsin, 1972.

THOMAS, E. A. C. Sufficient conditions for monotone hazard rate: An application to latency-probability curves. *Journal of Mathematical Psychology*, 1971, **8**, 303–332.

RECEIVED: December 30, 1974