JOURNAL OF VERBAL LEARNING AND VERBAL BEHAVIOR 17, 493–507 (1978)

Individual Differences in the Verification of Sentence–Picture Relationships

COLIN M. MACLEOD, EARL B. HUNT, AND NANCY N. MATHEWS

University of Washington

In a modification of the familiar sentence–picture comprehension task (Chase & Clark, 1972), 70 university undergraduates verified simple sentence–picture pairs. Two reaction times were collected on each trial: (a) comprehension time, the time to study a sentence of the form PLUS IS (NOT) ABOVE STAR, and (b) verification time, the time to verify whether a picture of the form $\overset{+}{*}$ was true with respect to the sentence. The verification reaction times of individual subjects were fit to the Carpenter and Just (1975) constituent comparison model and two groups of subjects were isolated. The larger group was well fit by the model, indicating that they adopted a linguistic strategy. The smaller group was poorly fit by the model; their reaction time pattern suggested use of a pictorial-spatial strategy. Psychometric measures confirmed a clear difference between the two groups in spatial ability but not in verbal ability. This difference was consistent with the hypothesized verification strategies; the subjects using the pictorial-spatial strategy demonstrated markedly higher spatial ability. These findings limit the generalizability of any linguistic comparison model by demonstrating that two quite different comprehension strategies are used consistently by different subjects. More important, the subject's choice of strategy is predictable from his psychometric measures of cognitive ability.

One of the basic tasks in language comprehension is deciding whether a linguistic statement truly describes our observations about the world. How do we do this? On logical grounds alone, we know that we must somehow form common representations of the linguistic and the nonlinguistic stimuli before this decision can be made. How are these representations formed and compared? This question is a fundamental one for psycholinguistics.

In response, psychologists have conducted extensive studies of the time people require to verify sentences about quite simple pictures. The basic paradigm was developed by Clark and Chase (1972). The subject first observes a simple sentence, such as PLUS IS ABOVE STAR, and then a picture, either $\overset{*}{+}$ or $\overset{+}{*}$. The task is to indicate, as rapidly as possible, whether the sentence is a true description of the picture. The chief independent variable is the linguistic complexity of the sentence. For example, typical sentences might be PLUS IS ABOVE STAR, STAR IS NOT ABOVE PLUS, or PLUS IS BELOW STAR. The dependent variable is reaction time, as error rates are held quite low.

Within this paradigm, there are four primary situations produced by combining affirmative or negative sentences (e.g., PLUS IS ABOVE STAR, PLUS IS NOT BELOW STAR) with pictures for which the sentences are either true or false propositions. Table 1

We are grateful to Kevin Quinn for his assistance in collecting the data, to Colene McKee for her assistance with the data analysis, and to Marcy Lansman and John Palmer for their helpful criticisms. This research was supported by National Institute of Mental Health grant MH-21795 and National Institute of Education grant G77-0012 to the University of Washington.

Requests for reprints may be sent to Colin M. MacLeod, Division of Life Sciences, Scarborough College, University of Toronto, West Hill, Ontario M1C 1A4, CANADA.

0022–5371/78/0175–0493$02.00/0

TABLE 1

THE SENTENCE–PICTURE STIMULUS PAIRS AS A FUNCTION OF TRIAL TYPE, HYPOTHETICAL REPRESENTATION, AND NUMBER OF CONSTITUENT COMPARISONS

Trial type	Sentence	Picture	Sentence representation	Picture representation	Number of constituent comparisons
True affirmative (TA)	STAR IS ABOVE PLUS PLUS IS BELOW STAR	* +	[AFF(STAR, TOP)]	(STAR, TOP)	K
False affirmative (FA)	PLUS IS ABOVE STAR STAR IS BELOW PLUS	* +	[AFF(PLUS, TOP)]	(STAR, TOP)	$K + 1$
True negative (TN)	PLUS IS NOT ABOVE STAR STAR IS NOT BELOW PLUS	* +	{NEG[AFF(PLUS, TOP)]}	(STAR, TOP)	$K + 5$
False negative (FN)	STAR IS NOT ABOVE PLUS PLUS IS NOT BELOW STAR	* +	{NEG[AFF(STAR, TOP)]}	(STAR, TOP)	$K + 4$

Note. The constituent comparison model (Carpenter & Just, 1975) predicts TA < FA < *FN* < TN.

illustrates these possible combinations. We shall refer to the four basic situations as True Affirmative (TA), False Affirmative (FA), True Negative (TN), and False Negative (FN) trials.

Carpenter and Just (1975) have presented a model for both the formation of representations and the comparison process in sentence verification. The model contains three assumptions:

(a) *Sentence representation.* Sentences are represented internally by logical propositions which are equivalent to the sentences. The propositions are a function of the surface structure of the sentence. Table 1 shows the propositional form assumed for each sentence.

(b) *Picture representation.* Pictures are represented internally by logical propositions equivalent to the affirmative statement which describes them.

(c) *Comparison process.* After both representations have been formed, they are compared, component by component, from the innermost to the outermost constituent. (Hence the name used by Carpenter and Just, the "constituent comparison model.") When a mismatch is detected, the two offending constituents are marked "resolved" and the comparison process begins anew. The process is terminated when all constituent comparisons are found either to result in agreement or to involve "resolved" components. At this point a response is output. The value of the response can be deduced by determining whether there have been an even or odd number of attempts to complete a comparison.

Let us call each sequence of comparisons a "scan." Each of the four trial types (TA, FA, TN, FN) will require a different number of scans. Carpenter and Just further assume that each scan requires a constant amount of time, which we arbitrarily set to one unit of time. We also make the simplifying assumption that, taken together, initial coding of the picture and response production require k units of time. Given these assumptions, Carpenter and Just's model predicts that the average amount of time required for each trial type will vary from k units for a TA trial to $k + 5$ units for a TN trial. Table 1 shows the number of units predicted for each of the four situations. The constituent comparison model, in effect, places each trial type at a unique point on an interval scale, and predicts that observed reaction time in sentence verification (which, it will be recalled, is measured from the onset of the picture) will be a linear function of this scale. Carpenter and Just (1975, Tables 4, 5, 7, & 8) reviewed a number of studies and argued that the linear model effectively captured a very large percentage of the variance in reaction time across conditions. Although the model has been criticized on both theoretical and empirical grounds (Catlin & Jones, 1976;

Tanenhaus, Carroll, & Bever, 1976), the data reported by Carpenter and Just are indeed impressive.

Carpenter and Just cautiously say that "the internal representation of a sentence is not necessarily linguistic in nature" (1975, p. 47), and refer to the internal representation as an abstract propositional form. In practice, though, the particular propositional form used for each sentence is a function of the linguistic structure of the sentence being represented, so it seems fair to argue that the Carpenter and Just model is one of a very wide class of models in which the linguistic form of a sentence as well as its logical interpretation influences its internal representation.

Most studies using the sentence verification task have gathered data from a small number of highly trained subjects. The psycholinguistic assertions that are made, however, are obviously intended to be assertions about how people represent linguistic statements in general. One would hope that these assertions are correct, for if they are, there is a single parameter of the model, the slope parameter, which is esentially a measure of how long it takes the subject to complete a single scan. This slope could then be used as a theoretically justified measure of an important process in language comprehension. In previous studies (Hunt, Frost, & Lunneborg, 1973; Hunt, Lunneborg, & Lewis, 1975), individuals with varying degrees of verbal skill, as measured by conventional psychometric tests of verbal ability, have been shown to differ in the time with which they do numerous "basic" tasks assumed to be essential in verbal comprehension (e.g., identifying the names of letters). Preliminary studies (cf. Hunt *et al.*, 1975) suggested that sentence verification times do covary with verbal ability, and indeed, that it might be possible to construct a "paper and pencil" test which would measure the process on an individual basis (Baddeley, 1968; Lansman & Hunt, Note 1). Because the Carpenter and Just model yields a single parameter which can be justified by a psycholinguistic model, we were particularly interested in knowing whether it could be used as a measure of comprehension in a battery of tests of language skills based upon an information processing theory. In order to answer this question, however, we needed data verifying the model using a large group of subjects. Our goal also required that we collect a number of psychometric measures on the subjects. We hoped in this way to obtain a fairly detailed picture of how differences in cognitive ability affected comprehension processes both quantitatively and qualitatively.

Method

Stimuli. The stimuli were the eight sentence-picture pairs shown in Table 1 together with another eight pairs in which only the order of + and * in the picture was reversed. That is, four binary dimensions, (STAR, PLUS), (IS, IS NOT), (ABOVE, BELOW), and ($\genfrac{}{}{0pt}{}{+}{*}$, $\genfrac{}{}{0pt}{}{*}{+}$) were combined to form the 16 possible different sentence-picture pairs.

Apparatus. Stimulus presentation and response collection were controlled by a NOVA 820 computer. The control system allowed up to six subjects to participate simultaneously and independently. Subjects were seated in individual sound-attenuating booths, each of which contained a response keyboard and a Tektronix 604 display scope for presenting the stimuli.

Subjects and psychometric measures. The subjects were 70 University of Washington undergraduates whose participation partially fulfilled a course requirement. Subjects were run in groups of one to four.

Three psychometric measures of ability (comprehension, verbal, and spatial) were available. Form A of the Nelson–Denny (1960) reading test was administered to all 70 subjects, yielding a comprehension score. This score was the number of correct answers to a series of multiple-choice questions following each of several passages in the test. Normally the comprehension section is terminated after

20 min. Instead, we allowed our subjects to finish the section, working under instructions to proceed as quickly as possible, without sacrificing accuracy. Mean completion time was 22.57 min (*SD* = 5.94).

In addition, some of the subjects made available their scores on the Washington Pre-College (WPC) test. The WPC is a group-administered scholastic aptitude test similar to the widely used Scholastic Aptitude Test (SAT). The test is taken by high school juniors in the state of Washington who are considering further education; thus, most subjects had taken the WPC test two to three years earlier. For 48 subjects, we had access to a composite verbal ability score; 46 of these subjects also had a spatial ability score in their files.

The WPC test is made up of several subtests. For our purposes, only the verbal composite and spatial ability measures are of interest. The verbal composite score is a weighted average of the vocabulary, English usage, spelling, and reading comprehension subtest scores. The spatial test requires the subject to visualize how a two-dimensional figure would look in three dimensions if folded along certain lines. More details of the WPC can be found in the Technical Manual (Note 2).

Procedure. After familiarizing subjects with the apparatus, partly by conducting a simple reaction time task, instructions for the verification task were given. These were:

> You are going to be asked to make judgments about whether a simple picture is true in relation to a sentence. (Two examples on index cards were shown and explained.) Here's how the task will work. First, you will see the sentence for as *long* as you need. For example, STAR IS ABOVE PLUS may appear. When you are ready for the picture, press either button. A half-sec later, a picture, either plus above star or star above plus, will appear. Your task is to indicate whether this picture is *true* with relation to the sentence you just read. If it *is*, press the TRUE button; if *not*, press the FALSE button. Then the next sentence will appear, and so on. What we are interested in is *how long* you spend in reading the sentence and on making your True–False judgment for the picture. You should try to go as *quickly* as you can, *without* making *errors*.

The feedback procedure and practice trials were then described, and the subjects were reminded that trial types were randomized. Subjects were also informed to use their left index fingers for FALSE responses and their right index fingers for TRUE responses. Finally, there was a brief review of the instructions, encouraging the subjects "to read the sentence and to make your judgment as quickly as you can, avoiding errors."

After the instructions, subjects did two blocks of 16 practice trials. Within each block, each of the 16 sentence–picture pairs was presented once, in a random order. Subjects were given the opportunity to ask procedural questions after the instructions and after each practice block. After practice, there were two blocks of 64 experimental trials, with a short break between blocks. Each experimental block contained four repetitions of the 16 stimulus pairs (i.e., 16 examples of each trial type in Table 1) with repetitions distributed randomly throughout the block.

On each trial, a warning dot appeared for 500 msec, followed by the stimulus sentence, which was presented horizontally at the center of the screen. When ready, the subject pressed either key and the picture replaced the sentence after 500 msec. The first reaction time (Comprehension RT) on a given trial was the time from sentence onset to the initial key press. The second reaction time (Verification RT) was the time from picture onset until the subject pressed the TRUE or FALSE key.

Immediately after the subject's response, a 500 msec feedback message was displayed. If the subject made an error on the trial, the word WRONG was displayed. If the subject was correct, the word RIGHT was displayed together with the Verification RT for that trial. Subjects were not informed of their Comprehension RT at any point during the experiment. The time between offset of feedback and onset of the next warning dot was 500 msec.

Results and Discussion

Outliers and Reliability

The mean reaction time and standard deviation of reaction times for every trial type were calculated for each individual. Data were analyzed only for those trials on which the subject was correct and on which the Comprehension and Verification reaction times were within three standard deviations of their respective means. This criterion eliminated only 4% of the trials, those with extremely short or long reaction times (i.e., greater than 5 sec or less than 200 msec).

Because part of our interest centered on the use of the sentence verification task in individual differences studies, it was necessary to show that this task does place individuals in a reliable ordering relative to each other. This can be established by calculating split-half reliabilities (odd vs even trials, separately for affirmative and negative trial types for each subject) and then applying the Spearman–Brown formula to estimate total task reliability. These reliabilities were .97 (affirmative) and .98 (negative) for Comprehension RTs, and .99 (affirmative) and .97 (negative) for Verification RTs.

Entire Group Performance

Table 2 displays the mean RTs, averaged over subjects, as a function of sentence type (Comprehension RT: Affirmative vs Negative) and sentence–picture relationship (Verification RT: TA, FA, TN, and FN). Below each Verification RT is its respective error rate; the mean error rate of 9.5% is comparable to that in other studies of sentence–picture verification even though our subjects had rather less practice than is typically the case.

Comprehension RT. As is evident from the values presented in Table 2, mean Comprehension RT was significantly longer for negative sentences than for affirmative sentences [$F(1, 69) = 151.7$, $MS_e = 91{,}151$, $p < .001$]. This is consistent with the frequently reported finding that the insertion of a negative term increases sentence processing time (e.g., Gough, 1965; Wason & Jones, 1963).

Verification RT. A two-way ANOVA was conducted on Verification RTs, with True–False and Affirmative–Negative as factors. Both main effects were highly significant, with False responses requiring longer than True responses [$F(1, 69) = 29.9$, $MS_e = 28{,}943$, $p < .001$] and Negative responses requiring longer than Affirmative responses [$F(1, 69) = 75.4$, $MS_e = 90{,}008$, $p < .001$]. The interaction was also significant [$F(1, 69) = 22.5$, $MS_e = 24{,}218$, $p < .001$], demonstrating a

TABLE 2

Mean Comprehension RTs, Verification RTs, and Error Rates as a Function of Trial Type for All 70 Subjects

	Affirmative		Negative	
RT	True	False	True	False
Comprehension	1575		2203	
Verification	773	972	1172	1195
Percentage errors	7.0	7.9	13.2	9.8

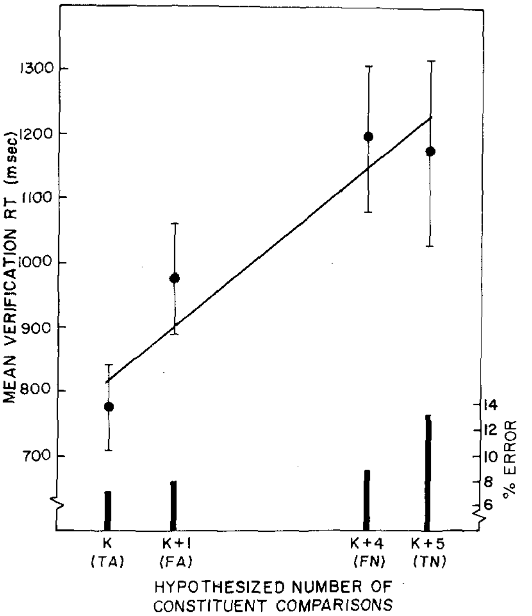

FIG. 1. Mean Picture RT as a function of hypothesized number of constituent comparisons (trial type is shown in parentheses). Also included are the 95% confidence intervals and the best-fitting straight line (intercept at 749 msec, slope of 79.7 msec per comparison).

larger effect of negation on True than on False trials.

Figure 1 shows the four Verification RT means arrayed in the order predicted by the Carpenter and Just (1975) constituent comparison model. The predicted linear increase of Verification RT over the four trial types (TA, FA, FN, TN) accounts for a highly significant 89.4% of the variance [$F(1, 207) = 153.5$, $MS_e = 47{,}723$, $p < .001$]. Although this is consistent with the model, it is not as high as might be expected. Furthermore, the residual 10.6% of the variance is also significant [$F(2, 207) = 9.2, p < .001$]. This significant residual was unexpected in light of the usual 95–99% of the variance accounted for by the model over a wide set of data (see Tables 4, 5, 7, & 8 in Carpenter & Just, 1975). In particular, although there is an interaction between the Affirmative–Negative and True–False factors, our negative conditions are ordered TN < FN, not FN < TN as the model predicts. We will offer an account of this reduction in goodness of fit in our subsequent examination of individual differences.

Markedness effects. Separate analyses of the effect of marked (BELOW) vs unmarked (ABOVE) prepositions were conducted on both Comprehension RT and Verification RT. Marked sentences ($\overline{RT} = 1918$ msec) took significantly longer to comprehend than did unmarked sentences ($\overline{RT} = 1862$ msec) [$F(1, 69) = 8.2$, $MS_e = 26{,}487$, $p < .01$]. In addition, pictures following marked sentences ($\overline{RT} = 1059$ msec) took significantly longer to verify then did pictures following unmarked sentences ($\overline{RT} = 1001$ msec) [$F(1, 69) = 16.7$, $MS_e = 28{,}830$, $p < .001$]. These differences are consistent with markedness effects reported in the rest of the sentence–picture verification literature.

Error analysis. A two-way ANOVA was conducted on the number of incorrect responses as a function of True–False and Affirmative–Negative. There were significantly more errors on negative trials than on affirmative trials [$F(1, 69) = 23.1$, $MS_e = 5.12, p < .001$]. Although there was a trend for more errors on True trials than on False trials, this effect was not significant [$F(1, 69) = 2.0$, $MS_e = 5.29$, $p = .17$]. However, the interaction was significant [$F(1, 69) = 7.4$, $MS_e = 4.43$, $p < .01$], indicating that negation affected True responses more than False responses. Overall, the error pattern is similar to the correct RT pattern and error rates for the four conditions are highly correlated with RTs ($r = .81$). This reduces any concerns regarding a speed-accuracy tradeoff, since both errors and RTs increase with the number of hypothesized operations in the constituent comparison model.

Interpretation and summary. Carpenter and Just's (1975) constituent comparison model offers a reasonably good fit to the Verification RTs for the entire group of 70 subjects. However, although the linear trend dominates, there is significant nonlinearity as well. Indeed,

TABLE 3

CORRELATIONS OF WASHINGTON PRE-COLLEGE TEST PERFORMANCE WITH COMPREHENSION RTS AND VERIFICATION RTS IN THE VERIFICATION TASK

	Comprehension RT			Verification RT			
WPC test	Affirmative	Negative	Slope	TA	FA	FN	TN
Verbal	–.24	–.16	–.33	–.49	–.58	–.49	–.47
Spatial	.07	.10	–.54	–.46	–.55	–.60	–.57

Note. Correlations with verbal ability are based on 48 subjects; those with spatial ability include only 46 of the same subjects. Negative correlations indicate that higher psychometric scores are related to faster RTs.

this is quite evident in the reversal of FNs and TNs with respect to the model.

Carpenter and Just handle the TN–FN reversal when it occurs with the extra assumption of *recoding*—subjects may change a negative representation to an affirmative one before performing the comparisons. In fact, Carpenter and Just even claim that recoding is encouraged by "a delay between the presentation of the sentence and the second source of information" (p. 66; see also Carpenter, 1973; Trabasso, 1972), a situation to which our two-RT method may be analogous. Furthermore, in a separate study in our laboratory using simultaneous sentence-picture displays, Lansman and Hunt (Note 1) have found that about half of their subjects show the predicted ordering (FN < TN) while the other half show the reverse ordering (TN < FN). This parallels our findings. Perhaps, then, some of our subjects were recoding negative sentences before going on to the verification stage. We shall examine this hypothesis below, and offer an alternative theoretical account that does not rely on recoding.

Psychometric measures. Table 3 summarizes the correlations of the WPC verbal and spatial ability tests with the Comprehension RTs and Verification RTs in the sentence-picture task. Because the mean scores in both of the WPC tests (Verbal: $\bar{X} = 54.3$, $SD = 8.06$; Spatial: $\bar{X} = 53.2$, $SD = 10.03$) resemble the population values ($\bar{X} = 50$, $SD = 10$), we may be confident that we are looking at a representative sample. We first note that the verbal and spatial tests are themselves significantly correlated ($r = .59$, $n = 46$, $p < .001$). Although neither test predicts Comprehension RT very well (none of the correlations is significant), both are good predictors of Verification RT and of the slope parameter (all correlations $p < .01$). The relationship between verbal ability and the verification task measures is further supported by results from a different psychometric test. The Nelson–Denny comprehension scores were also negatively correlated with the Verification RTs ($r = -.41, p < .001$) and the slope ($r = -.31, p < .005$). Thus, it is reasonable to assume that the sentence-picture verification task is tapping some of the same skills measured by traditional psychometric techniques.

The magnitude and consistency of both sets of Verification RT correlations shown in Table 3 is striking, particularly since spatial ability predicts at least as well as verbal ability in all cases. The predictive power of spatial ability had not been anticipated because the existing sentence-picture literature relies on a linguistic (i.e., verbal) account of performance in the task. This finding intensified our interest in examining individual differences in the task.

Patterns of Individual Differences

Although the overall pattern of results for the entire group of 70 subjects is largely consistent with the predictions of the constituent comparison model, there are exceptions. We have already indicated the high correlations of task performance with spatial ability, surprising in view of the linguistic emphasis of the model. Also, the significant nonlinearity in the Verification RTs is difficult to reconcile with the model. We will now examine the verification data from the standpoint of differences between individuals in the extent to which their data fit the model. Our intention is to demonstrate, using psychometric measures as support, that a single model is inadequate for capturing the intersubject variability in sentence–picture verification.

Individual fits to the model. The first step in examining the data of individual subjects was to determine how well each subject's data were fit by the constituent comparison model. For each individual, a correlation was computed between the four Verification RTs and their predicted number of comparisons. Although the median correlation was quite high (r = .82), the range was exceedingly wide (from .998 to −.877) certainly not consistent with the model's predictions.

The next step was to break down the subjects into those who were well fit and those who were poorly fit by the model. To accomplish this, we rank-ordered the subjects in terms of their correlations with the model's predictions and then split the subjects into three groups by applying a variant of Fisher's clustering algorithm for one-dimensional data (Hartigan, 1975). We first divided the sample into two subgroups such that a t-test of the difference between mean correlations for the two subgroups was maximized. This identified a group of 16 subjects who were "poorly fit" by the constituent comparison model. The same procedure was then applied to split the larger group of 54 subjects into two further subgroups, 43 subjects who were "well fit" by the model, and 11 subjects of "intermediate fit," whose data were not clearly interpretable. Table 4 presents the statistics describing these sub-groups relative to the constituent comparison model. Our subsequent discussion will focus only on the well-fit and poorly-fit groups.

Verification RTs. The analysis of correlations simply tells us that particular subjects' Verification RT data do or do not bear a linear relationship to the predictions of the constituent comparison model. Figure 2 is a detailed illustration of the form of the relationship in the well-fit group and the lack of it in the poorly-fit group. The data from the well-fit group are almost perfectly fit by the model. The linear trend accounts for 97.8% of the variance across conditions; the residual 2.2% is nonsignificant. This, of course, is a noninformative statement, as the well-fit group was selected so that their data would fit the predictions of the model. The picture for the poorly-fit group is of considerably greater interest. Figure 2 suggests that the only factor affecting Verification RTs in this group is the True-False distinction. This was confirmed by an analysis of variance of Verification RTs in the poorly-fit group; the only significant effect was the True–False distinction [$F(1, 15) = 20.5$, $MS_e = 22{,}638$, $p < .001$]. Both the Affirmative–Negative effect and the interaction yielded F ratios of less than 1. Thus,

TABLE 4

INDIVIDUAL SUBJECT CORRELATIONS TO THE CONSTITUENT COMPARISON MODEL AS A FUNCTION OF GOODNESS OF FIT

Group	Number of subjects	Range of correlations	Median correlation
Well-fit	43	.679 to .998	.934
Intermediate-fit	11	.378 to .603	.467
Poorly-fit	16	−.877 to .285	.009

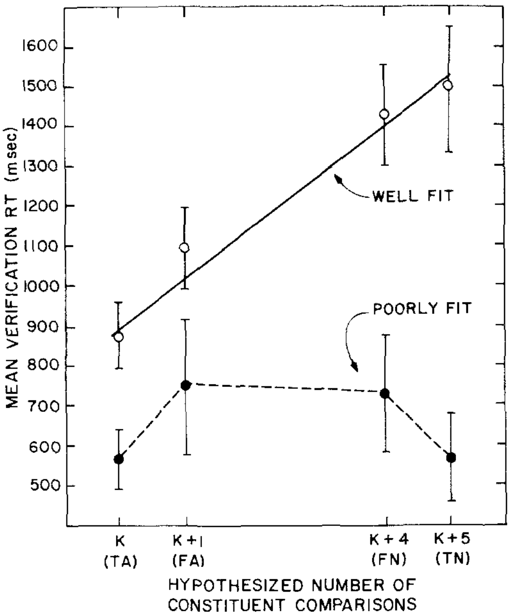

FIG. 2. Mean Picture RT as a function of hypothesized number of constituent comparisons. The curve parameter is group, Well-Fit vs Poorly-Fit to the Carpenter and Just (1975) model. Also included are the 95% confidence intervals, and the best-fitting straight line for the Well-Fit subjects only (intercept at 797 msec, slope of 121 msec per comparison).

these "linguistic" effects have disappeared as an influence upon verification times in the poorly-fit group.

Markedness effects. But what of the other linguistic effect, that of markedness? Table 5 displays the relevant data. The markedness effect is robust in the well-fit group for both the Comprehension RTs [$F(1, 42) = 5.0$, $MS_e = 23{,}176$, $p < .05$] and the Verification RTs [$F(1, 42) = 23.7$, $MS_e = 35{,}671$, $p < .001$]. This is consistent with a linguistic model. However, for the poorly fit group, the picture is quite different. The marginally significant effect of markedness in the Comprehension RTs [$F(1, 15) = 3.9$, $MS_e = 48{,}239$, $p = .07$] is absent in the Verification RTs [$F(1, 15) = 1.2$, $MS_e = 1434$]. Apparently, these subjects have eliminated the ABOVE–BELOW distinction by the time they reach the verification stage. Taken together with the absence of the Affirmative–Negative effect and the interaction noted above, the poorly-fit group is poorly fit by *any* linguistic model, not just by the Carpenter and Just model.

Before further discussion of the RT data on the two groups, it should be noted that the error patterns again correspond with the RT patterns. Errors were positively correlated with RTs in both the well-fit group ($r = .88$) and the poorly-fit group ($r = .72$). This obviates concerns regarding a speed-accuracy tradeoff in either group's data.

A comparison of alternative general models. We will next consider the implications of two distinct models of sentence verification times, a general linguistic model and a general pictorial model. We shall argue that these models provide qualitatively different views of what subjects are doing in the sentence verification task, that the two models are required to account for the results which we have obtained, and that it is possible to predict from subject characteristics which models should be applied to which subjects.

TABLE 5

MEAN COMPREHENSION RT AND VERIFICATION RT FOR MARKED VS UNMARKED TRIALS AS A FUNCTION OF WELL-FIT VS POORLY-FIT GROUP IDENTIFICATION

Group	Comprehension RT		Verification RT	
	Unmarked (Above)	Marked (Below)	Unmarked (Above)	Marked (Below)
Well-fit	1633	1685	1173	1272
Poorly-fit	2529	2637	649	656

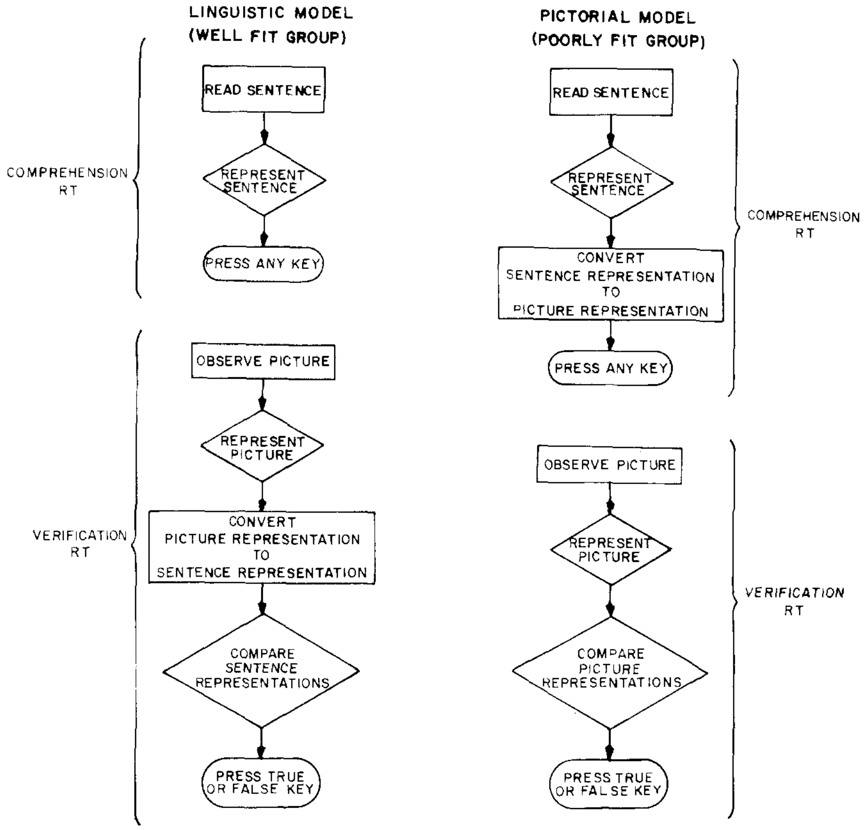

FIG. 3. Sentence–picture verification models for the two-RT procedure. On the left is the linguistic model; on the right is the pictorial-spatial model.

Figure 3 presents both models in a flowchart notation for describing sentence verification. We first consider the general linguistic model. Referring to the left side of Fig. 3, the model assumes that:

(1) When the sentence is presented (and read), a "linguistically based" propositional form of the sentence is developed.

(2) The subject indicates that the sentence has been understood. (Steps 1 and 2 constitute the Comprehension (sentence) RT.)

(3) The picture is presented and an internal visual representation of the picture is formed.

(4) The internal representation of the picture is converted to a propositional form equivalent to that of a true assertion about the sentence.

(5) The two propositional forms are compared. The time taken for this comparison will be partly a function of the complexity of the proposition formed at Step 1, as the complexity of the picture representation is constant.

(6) A response is emitted. (Steps 3–6 constitute the Verification (picture) RT.)

The gist of the general linguistic model is that the internal representations of both sentence and picture are converted to propositional form. An obvious alternative is that the propositional form of the sentence is used to generate an "expected picture" representation, and that when the visual representation of the picture is formed, it is compared directly to the "expected picture" representation. This sequence is shown in the flow chart for a general pictorial model on the right-hand side of Fig. 3. The pictorial model differs from the sentence model in that:

(a) The step of converting the sentence representation to a picture representation is added *into* the Comprehension (sentence) RT.

(b) The step of converting the internal picture representation into a sentence-based proposition is eliminated from the Verification (picture) RT.

(c) The comparison process which takes place during verification will no longer be a function of the linguistic structure of the sentence, since this information will have been removed from the internal representation when the "expected picture" representation was generated.

The idea of a pictorial representation is not original with us (for related discussion, see Clark & Chase, 1972). In fact, Tversky (1975) has shown that a pictorial representation is a more likely explanation for the data when there is a separation between sentence and picture presentation, while the linguistic model appears to apply when sentence and picture are presented simultaneously. Yet Carpenter and Just (1975, Experiment 2) present data indicating that the linguistic model remains appropriate if the sentence is presented for a fixed time (2 sec) immediately prior to picture presentation. We want to consider the possibility that in our situation, in which the subject controls the time the sentence is presented, different subjects will choose different processes.[1] Clearly, this assertion is compatible with the different Verification RT data for our two groups. However, the argument is post hoc, since the well-fit group was chosen so that a sentence structure effect was inevitable.

An independent test of the two models as descriptions of the two groups' data is possible. Consider the pattern of sentence comprehension reaction times as a function of group membership, and the difference between sentence-comprehension and verification reaction time patterns in the two groups. These comparisons are independent of the operations used to define the groups, because those operations considered data based only on the Verification RTs. We can also consider the difference between group mean Verification RTs. This difference will be independent of group definitions, since the groups were defined using correlations of Verification RTs with the Carpenter and Just metric, and correlations are independent of mean values of the variables being correlated.

Assuming that there is no correlation between motor response processes and model use, which seems a reasonable assumption, individuals who follow the pictorial model should take longer in sentence comprehension than individuals who follow the linguistic model. This is so because the pictorial-model subjects must execute the additional step of converting a linguistically based proposition into an expected pictorial representation prior to indicating sentence comprehension. On the other hand, during the verification stage, the pictorial-model subjects would not have to convert the initial visual representation of the picture into a propositional form. Therefore, pictorial-model subjects should be faster in the verification stage. Table 6 presents the mean comprehension and verification reaction times for the well-fit and the poorly-fit groups. The relations are as predicted on the assumption

TABLE 6

MEAN OVERALL COMPREHENSION RT AND VERIFICATION RT FOR THE WELL-FIT AND POORLY-FIT GROUPS

Group	Comprehension	Verification
Well-fit ($n = 43$)	1652	1210
Poorly-fit ($n = 16$)	2579	651

[1] The roles of instructions to the subject and of task structure are clearly crucial, as Glushko and Cooper (in press) have recently shown. We simply point out that subject-pacing of verification tasks is probably most conducive to individual strategy choice.

TABLE 7

CORRELATIONS OF PSYCHOMETRIC SCORES WITH MEAN VERIFICATION RT

Group	Nelson–Denny comprehension	WPC verbal	WPC spatial
Well-fit	–.47*	–.52*	–.32
Poorly-fit	–.03	–.33	–.68*

Note. Those correlations marked with an asterisk are significant beyond $p < .01$.

that the well-fit group follows a linguistic model and the poorly-fit group follows a pictorial model.

Psychometric scores. In our final analysis we will use psychometric scores to further our argument that the different groups were using different strategies, and to offer some evidence indicating that it is possible to predict which subjects will use which strategies. Except for the comprehension scores, these analyses will be based upon the 39 subjects (27 in the well-fit group and 12 in the poorly-fit group) on whom WPC measures were available.

The pictorial model assumes that during the verification stage subjects engage in comparison of visual images, something psychometricians would refer to as a task involving spatial abilities. If this is so, we would expect spatial ability measures to be negatively correlated with Verification RTs in the poorly-fit group (i.e., subjects with better spatial ability should have faster RTs). On the other hand, according to the linguistic model, verification time is determined by the same process of propositional comparison that would determine the comparison between two verbal statements, so we would expect verbal ability and comprehension performance to be (again, negatively) correlated with Verification RT.

The relevant data for mean Verification RTs are presented in Table 7. Clearly, the relative magnitude of the correlations is as expected. The problem is complicated, however, by the fact that there are substantial (although statistically nonsignificant) correlations between verification time and spatial ability in the well-fit group, and between verification time and verbal ability in the poorly-fit group. This would be expected, on statistical grounds alone, if there is a positive correlation between measures of verbal and spatial ability, as indeed there is in our sample ($r = .59$). Therefore, a somewhat better picture of the relative relationship between the WPC measures and task performance is obtained by calculating partial correlations, in which the correlations between reaction times and verbal ability are computed with spatial ability "held constant" and vice versa. The result of this analysis is shown in Table 8. It even more strikingly supports our claim that different models are needed for our two groups of subjects.

It is also interesting to examine how each group's psychometric test scores correlated with the slopes and intercepts for Verification RT derived from the Carpenter and Just model. These correlations are shown in Table 9. In the well-fit group, both parameters are related to verbal ability but not to spatial ability. This is consistent with the idea that the intercept measures the time to construct a linguistic representation and the slope

TABLE 8

PARTIAL CORRELATIONS OF WPC VERBAL AND SPATIAL ABILITY WITH VERIFICATION RT

	Verbal ability Spatial ability = partial correlate	Spatial ability Verbal ability = partial correlate
Well-fit group	–.44*	.07
Poorly-fit group	–.05	–.64*

Note. Those correlations marked with an asterisk are significant beyond $p < .01$.

TABLE 9

CORRELATIONS OF WPC VERBAL AND SPATIAL ABILITY WITH MODEL-BASED SLOPES AND INTERCEPTS OF THE VERIFICATION RTS FOR THE WELL-FIT AND POORLY-FIT GROUPS

Group	Verbal ability	Spatial ability
Well-fit		
Slope	−.32*	−.26
Intercept	−.48*	−.19
Poorly-fit		
Slope	.10	−.11
Intercept	−.31	−.66*

Note. Those correlations marked with an asterisk are significant beyond $p < .05$. The slope for the poorly-fit group is based only on the True–False difference.

measures the time to compare the two linguistic representations. In the poorly-fit group, the only significant correlation is between spatial ability and the intercept, which indicates the time to construct a spatial representation. A "slope" based only on the True–False difference in the poorly-fit group failed to show any reliable relationship.

It is worth pointing out that the observed correlational pattern in the poorly-fit group is not predicted by the recoding model mentioned earlier. The recoding assumption is a linguistic one, not a spatial one. Thus, although the notion of recoding appeared consistent with the reaction time data, it is not consistent with the psychometric data. The pictorial model, which does predict the observed correlational pattern, offers a more parsimonious account than does any verbally-based model.

There is one further, tangential piece of evidence to support the contention that the two groups were using different strategies. Numerous studies have shown that men, in general, have higher spatial ability than women (cf. Maccoby & Jacklin, 1975). On the basis of this, we would expect a correlation between Verification RT and sex in the poorly-fit group, but not in the well-fit group, since only in the poorly-fit group should spatial ability be a factor in performance. Such a pattern was apparent. The correlation between sex and Verification RT was .55 ($p < .05$, 69% men) in the poorly-fit group, and .04 (56% men) in the well-fit group.

Psychometric characteristics of strategy users. The analyses just presented all speak to the question of what strategy is being used within each group. A somewhat different question is "Do people select strategies appropriate to their individual talents?" Suppose that they do. We would then expect to find at least one of the two following statements to be true: (a) The well-fit group should have higher verbal ability, or (b) the poorly-fit group should have higher spatial ability. Naturally, the statements are not mutually exclusive. Verifying either one of them would be evidence of a certain amount of "metacognition," since people would be selecting strategies in accordance with their own abilities.

In our data, the question can be answered only for those subjects who made their WPC scores available to us. If these scores do differentiate the groups, the finding would be strong evidence for a stable bias toward strategy selection. We should point out that our subjects were predominantly university freshmen and sophomores, and that the WPC test is administered to high school juniors. Thus, a minimum of two years intervened between taking the WPC and performing the sentence–picture verification task.

Table 10 presents the WPC scores for the

TABLE 10

MEAN WPC VERBAL AND SPATIAL ABILITY SCORES FOR THE WELL-FIT AND POORLY-FIT GROUPS

Group	Verbal ability	Spatial ability
Well-fit	53 (1.64)	49 (3.80)
Poorly-fit	56 (3.03)	60 (3.88)

Note. The standard error for each mean is shown in parentheses.

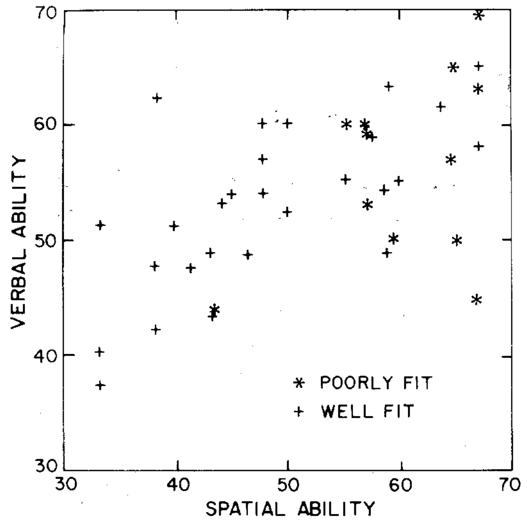

FIG. 4. Strategy choice as a joint function of WPC Verbal ability and Spatial ability.

two groups. Mean verbal ability and spatial ability scores, collapsed over groups, were not significantly different ($F < 1$). Collapsed over spatial and verbal ability, the two groups did differ significantly [$F(1, 37) = 7.8$, $MS_e = 111, p < .001$]. However, this group difference was qualified by a significant interaction of group with ability [$F(1, 37) = 9.3$, $MS_e = 27$, $p < .01$]. This interaction reflects the fact that the two groups did not differ on verbal ability scores, but that the poorly-fit group had markedly higher spatial scores. The effect of this difference in ability is shown in Fig. 4, which locates each subject in the plane defined by verbal and spatial ability scores, and also shows the strategy selected. The pictorial strategy was virtually never selected by subjects with spatial ability scores below 55. Above this level of spatial ability, the pictorial strategy was as common as the linguistic strategy.

CONCLUSION

Discussion of the appropriate models for psycholinguistic tasks is usually couched in general terms (i.e., "What models apply to people?"). Our results can be seen as a reminder that this approach is too simplistic. The same ostensibly linguistic task can be approached in radically different ways by different people. Our results should not be viewed as "disconfirmation" either of the general results on sentence verification or even of the specific model proposed by Carpenter and Just. Indeed, our results for the well-fit group can be viewed as strong support for that model. The subjects used in many of the relevant experiments have been drawn from the student bodies of universities such as Stanford and Carnegie–Mellon, institutions which follow restricted admissions policies. The types of processes observed within such a restricted range of abilities as is found in these populations may be quite unrepresentative of the problem-solving processes encountered in the general population. We point out that this remark applies not only to sentence–picture verification tasks; most studies of sentence verification and many other paradigms in cognitive psychology typically use intensively trained subjects. Intensive training may indeed have the effect of producing stable performance after the subjects discover and become proficient at a "most efficient" strategy for the laboratory task. Extra-laboratory generalization then becomes a problem (cf. Neisser, 1976).

The observation that untrained subjects will attack a task with a variety of strategies is neither new nor particularly interesting. However, our data indicate that strategy choice is a predictable function of subject abilities as measured by psychometric tests which, in the case of the subjects in our sample, were taken a minimum of two years prior to entering our laboratory. Furthermore, as the spatial ability scores indicate, it appears that strategy choice was, on the average, based on a rational estimate of the subject's own capabilities. While our results are discouraging for those who might wish to develop a single information processing theory of intelligence based upon parameter estimation, they are encouraging for those who pursue the more realistic goal of developing ways of identifying people who characteristically use certain information processing strategies, and then evaluating how well they use them.

References

Baddeley, A. D. A 3-minute reasoning test based on grammatical transformation. *Psychonomic Science*, 1968, **10**, 341–342.

Carpenter, P. A. Extracting information from counterfactual clauses. *Journal of Verbal Learning and Verbal Behavior*, 1973, **12**, 512–521.

Carpenter, P. A., & Just, M. A. Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review*, 1975, **82**, 45–73.

Catlin, J., & Jones, N. K. Verifying affirmative and negative sentences. *Psychological Review*, 1976, **83**, 497–501.

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

Clark, H. H., & Chase, W. G. On the process of comparing sentences against pictures. *Cognitive Psychology*, 1972, **3**, 472–517.

Glushko, R. J., & Cooper, L. A. Spatial comprehension and comparison processes in verification tasks. *Cognitive Psychology*, 1978, **10**, 391–421.

Gough, P. B. Grammatical transformations and speed of understanding. *Journal of Verbal Learning and Verbal Behavior*, 1965, **4**, 107–111.

Hartigan, J. *Clustering algorithms*. New York: Wiley, 1975.

Hunt, E., Frost, N., & Lunneborg, C. Individual differences in cognition: A new approach to intelligence. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 7). New York: Academic Press, 1973.

Hunt, E., Lunneborg, C., & Lewis, J. What does it mean to be high verbal? *Cognitive Psychology*, 1975, **7**, 194–227.

Maccoby, E. E., & Jacklin, C. N. *The psychology of sex differences*. Oxford: University Press, 1975.

Neisser, U. *Cognition and reality*. San Francisco: Freeman, 1976.

Nelson, M. J., & Denny, E. C. *The Nelson–Denny reading test*. Boston: Houghton Mifflin Co., 1960.

Tanenhaus, M. K., Carroll, J. M., & Bever, T. G. Sentence-picture verification models as theories of sentence comprehension: A critique of Carpenter and Just. *Psychological Review*, 1976, **83**, 310–317.

Trabasso, T. Mental operations in language comprehension. In J. B. Carroll & R. O. Freedle (Eds.), *Language comprehension and the acquisition of knowledge*. Washington, DC: Winston, 1972.

Tversky, B. Pictorial encoding of sentences in sentence–picture verification. *Quarterly Journal of Experimental Psychology*, 1975, **27**, 405–410.

Wason, P. C., & Jones, S. Negatives: Denotation and connotation. *British Journal of Psychology*, 1963, **54**, 299–307.

Reference Notes

1. Lansman, M., & Hunt, E. *Group testing procedures for measuring information processing variables*. Paper presented at the annual meeting of the Western Psychological Association, Seattle, April 1977.
2. University of Washington Assessment Center. *Technical manual: Washington pre-college test battery*. Seattle, 1977.

(Received October 21, 1977)