Component Latent Trait Models for Paragraph Comprehension Tests

Susan E. Embretson
University of Kansas

C. Douglas Wetzel
Navy Personnel Research and Development Center, San Diego

The cognitive characteristics of paragraph comprehension items were studied by comparing models that deal with two general processing stages: text representation and response decision. The models that were compared included the propositional structure of the text (Kintsch & van Dijk, 1978), various counts of surface structure variables and word frequency (Drum et al., 1981), a taxonomy of levels of text questions (Anderson, 1972), and some new models that combine features of these models. Calibrations from the linear logistic latent trait model allowed evaluation of the impact of the cognitive variables on item responses. The results indicate that successful prediction of item difficulty is obtained from models with wide representation of both text and decision processing. This suggests that items can be screened for processing difficulty prior to being administered to examinees. However, the results also have important implications for test validity in that the two processing stages involve two different ability dimensions.

Multiple-choice paragraph comprehension items are widely used in tests of verbal ability and educational achievement. The typical paragraph comprehension test consists of a short paragraph followed by a multiple-choice question about what has been read. Successful performance is affected by numerous sources of difficulty that are related to the way the paragraph text was constructed and the type of question asked about the text. These sources of difficulty are generally uncontrolled, unquantified, and dependent upon the artistry of the test item writer.

This paper demonstrates a method by which many sources of difficulty in paragraph comprehension test items may be quantified; the resulting indices are potentially useful for selection and banking of items by their cognitive characteristics. Quantifying the sources of item difficulty is important not only for understanding construct validity, but also for designing tests to reflect intended constructs (Embretson, 1983b, 1985). To implement the method, a cognitive processing model of the multiple-choice paragraph comprehension item is developed that specifies the sources of cognitive complexity.

Previous work with geometric analogy items (Whitely & Schneider, 1981) may serve as an easily understood example of this technique. Physical characteristics of the related pair of figures in the stem can be used to predict item difficulty. The relationship between the related pair is specified by the number of transformations that are required to change the first figure into the second, such as a change in shape or orientation of elements in the figures. In general, difficulty increases as the number of transformations increases. The enormously more complex situation of paragraph comprehension test items, with their linguistic and logical relations, can be approached using the same method. Fortunately, there are numerous theoretical ac-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 11, No. 2, June 1987, pp. 175–193

0146-6216/87/020175-19$2.20

counts of reading comprehension and question answering to guide this endeavor.

A few studies have attempted to understand the features of paragraph comprehension items that underlie item difficulty. Drum, Calfee, and Cook (1981) used stepwise regression to model the difficulty of paragraph comprehension items from the surface structure of the text, question stem, and response alternatives. The variables included number of content words, number of new content words, plausibility, word frequency, actual information, the requirement of outside knowledge, and sentence length. Although Drum et al. did not postulate a cognitive model of processing, they obtained good prediction of item difficulty for several reading tests composed of paragraph comprehension items.

From another perspective, Anderson (1972) has provided a convenient taxonomy of questions to test comprehension. His taxonomy can be seen as a dimension referring to the extent to which a transformation exists between the text of the paragraph and the choice alternatives. Anderson's six levels of questions are

1. Verbatim questions, in which a statement in the same form as the text is given as an alternative for verification;
2. Transformed verbatim questions, in which the same basic words are used as in the text, but the sentences or phrases are rearranged (e.g., "The boy hit the ball" becomes "By whom was the ball hit?");
3. Paraphrase questions, in which the question has the same meaning as a sentence in the text, but different words are used;
4. Transformed paraphrase questions, in which neither the wording nor the phrase order in the question is the same as in the text;
5. Alternative choices that are particular instances of a superordinate term in the question stem (i.e., deduction); and
6. Questions with particular instances in the question stem and alternative choices consisting of superordinate or gist statements (i.e., induction).

The cognitive characteristics of test items can be calibrated in the context of a latent trait model. The linear logistic latent trait model (LLTM; Fischer, 1973) has been used to calibrate cognitive item parameters and to test hypotheses about the cognitive components that underlie item solving. The LLTM is one of several component latent trait models (CLTM; Embretson, 1984) that link item responses to cognitive theory by predicting item difficulty from a mathematical model of response processing. The results of CLTM testing have direct implications for construct validity (Embretson, 1983b) in that CLTM elaborates predictive weights for the cognitive components that determine item difficulty. Furthermore, it is possible to use CLTM parameters to select items that have specified cognitive components.

Mitchell (1983) applied a LLTM to the paragraph comprehension items of the Armed Services Vocational Aptitude Battery (ASVAB), using some variables based on Kintsch and van Dijk's (1978) propositional analysis of text. Her model included propositional and argument density, word familiarity, number of words, and the number of inferences within the text and between the text and the correct answer. Her results were promising, but must be considered exploratory; only nine ASVAB items were available for the models and the parameters for the cognitive components were not tested for significance. Embretson and Wetzel (1984) used 30 items to replicate a portion of Mitchell's study and to examine some further new models reflecting the multiple-choice decision process. They found substantial improvement in prediction for a model that included more decision process variables. Their best model included word familiarity, one propositional analysis variable, and four decision process variables.

Although these diverse studies had positive findings, they do not constitute a sufficiently developed model to calibrate the cognitive characteristics of paragraph comprehension items. The purposes of the current study were (1) to propose a model of processing for multiple-choice comprehension items, (2) to compare the goodness of fit of the proposed model to previously studied predictors, (3) to calibrate the impact of the processing variables on

item difficulty, and (4) to examine the implications of the processing model for the construct validity of a paragraph comprehension test.

The proposed model contains two general processing stages: (1) a text representation process, in which the text is comprehended, and (2) a decision process, in which the question stem and the choice alternatives are compared for accuracy to the text. The text representation process is operationalized by Kintsch and van Dijk's (1978) propositional analysis of text, described below. The decision process is operationalized by some variables that have been shown to be important in modeling other multiple-choice items, such as verbal or geometric analogy items (Sternberg, 1977; Whitely & Schneider, 1981).

A Conceptual Model for Paragraph Comprehension Item Processing

Text Representation Processes

Figure 1 presents a very general conceptual model of the processing of paragraph comprehension items. The text representation process consists of two general events: lexical encoding (converting the visual stimuli of the text into a meaningful representation) and coherence processes (supplying information from memory about the preceding text as well as other facts and inferences to interpolate missing propositions that may make the sequence coherent).

The difficulty of lexical encoding was measured by word familiarity. In the current study, the Kucera-Francis (1967) index of word frequency was scored from the text question stem and alternatives. The Kucera-Francis index is the frequency with which a given word appears in samples of natural language text.

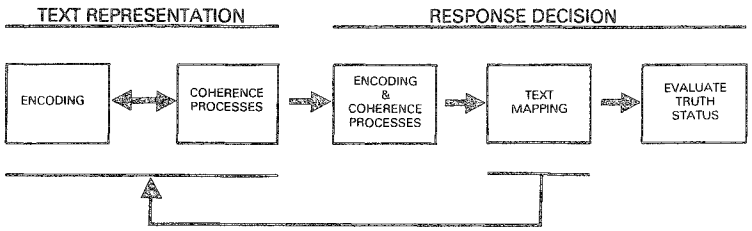
The difficulty of text processing can be measured by techniques for propositionalizing text developed by Kintsch and van Dijk (1978). They provided a theory-based approach to the problem of scoring the surface structure of textual material in terms of its meaning. Their theory assumes that the basic units of meaning are *propositions*, a more psychologically important feature of the text than its sentence surface structure. Propositions are composed of concepts, in which the first element is a *predicate* or relational concept, and the remaining one or more word concepts are *arguments*. The arguments are subjects and objects, while the predicate may consist of verbs, modifiers, or connectives. Thus the sentence "McGillacutty is a grey tabby cat" is propositionalized as:

P1. MODIFY: cat, tabby
P2. MODIFY: P1, grey
P3. ISA: McGillacutty, P2

The sentence contains two modifier propositions and one predicate proposition. Only unique arguments are coded, with previous propositions (e.g., P1) becoming the arguments of later propositions. Partially objective scoring systems to perform propositional analysis have been developed by Turner and Green (1978) and Bovair and Kieras (1981).

In the Kintsch and van Dijk model of text comprehension, the reader processes text in cycles, during which a network of coherent propositions is

Figure 1
General Information-Processing Model for Multiple-Choice Paragraph Comprehension Items



constructed. During a cycle, certain propositions are retained in a limited capacity short-term buffer for connection with input on the next cycle. An input is accepted as coherent with previous text if a connection is found between new propositions and those retained in the buffer. The reader may encounter difficulty during the cyclical construction of a coherent text base if no connection of the input to previously processed propositions is encountered, causing memory search, rereading, or generation of inferences to bridge gaps in the text.

Previous studies pertinent to Kintsch and van Dijk's work allow several expectations to be generated for the outcome of the current study. Kintsch and Keenan (1973) demonstrated that the number of propositions in a sentence, rather than the number of words, determines reading time. This finding validates the use of propositions rather than separate word counts for assessing text processing. If the number of words is held constant, texts with relatively more propositions should be more difficult. Thus, propositional density (the number of propositions divided by the number of words) should be related to item difficulty. Mitchell's (1983) results, however, were inconsistent with respect to the impact of propositional density on item difficulty. This inconsistency may be accounted for by separating propositions into the convenient typing provided in the Kintsch and van Dijk system. Propositional density was therefore scored separately for predicate, modifier, and connective propositions. Kintsch, Kozminsky, Streby, McKoon, and Keenan (1975) found that reading times were longer and recall was less for texts with many different arguments than for texts with fewer arguments. Propositions containing new arguments require an additional processing step on the part of the reader. (Recall probability also increased as a function of the number of repetitions of an argument in the text base.) Argument density is the number of unique arguments divided by the number of words, and the expectation was that higher density would be associated with less difficult items. Another finding from propositional analyses is that propositions are not equally difficult to remember, with superordinate propositions being recalled better than propositions that were structurally subordinate (Kintsch & Keenan, 1973; Kintsch et al., 1975). This effect was not operationalized in the current study because the paragraphs were relatively short.

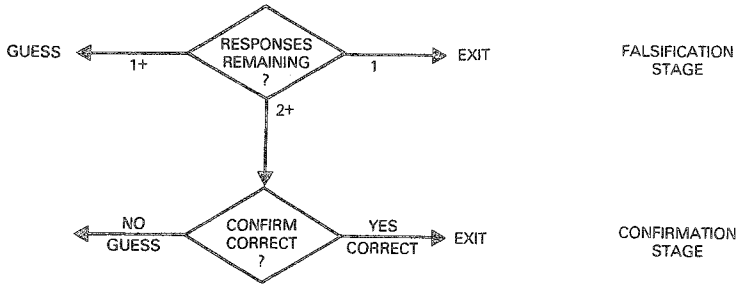
Decision Process

The decision process, outlined in Figure 1, is postulated to consist of three events: (1) encoding and coherence processes of converting the visual stimuli of the alternatives into a meaningful representation; (2) mapping the alternatives to the text, that is, finding the relevant text for evaluating each alternative; and (3) evaluating the truth status of the alternatives.

The first decision event, encoding, is operationalized in the same way as with the text processes. That is, the Kucera-Francis word frequency index and propositional analyses are applied to each alternative. The second decision event, mapping the alternatives to the text, requires locating the propositions that are relevant to falsifying or confirming the alternatives. If the relevant propositions are contained in a single short sentence, it is expected that text mapping would be easy. However, if the relevant propositions are spread throughout the text, or if the alternative concerns the "gist" of the paragraph, text mapping would be difficult. In the current study, text mapping difficulty was measured by the proportion of text over which the relevant propositions are located. The third decision event, evaluating the truth status of the alternatives, is shown in Figure 2. Consistent with studies of other item types (e.g., analogies), evaluating truth status is postulated to be a two-stage event (Pellegrino & Glaser, 1979; Whitely & Barnes, 1979).

The first evaluation stage is *falsification*, in which the examinee attempts to falsify as many alternatives as possible. In many multiple-choice items, the distractors are not wholly false or their truth status cannot be determined. In the initial componential models of decision processes (Sternberg, 1977), falsification was the method by which an alternative could be eliminated. However, in paragraph comprehension items, often no information in the text contributes to falsifying a distractor.

Figure 2
An Information-Processing Model for Evaluating the Response Alternatives



Thus, the alternative could possibly be true, but the paragraph is mute with respect to the assertions contained in the alternative. An item is difficult if few alternatives are falsifiable. In the current study, the difficulty of the falsification process was measured by 1.0 minus the probability that the correct answer could be selected by random guessing among the remaining non-falsified set.

The second evaluation stage is *confirmation*. This event was suggested as a decision process event for multiple-choice items in the Whitely and Barnes (1979) study of analogies. In confirmation, the remaining response alternatives are compared to the text to determine if the text confirms the response as correct. An alternative that is confirmed is selected as the correct answer. In the current study, a dichotomous variable was scored according to whether the text confirms the correct answer.

Two variables may interact with falsification and confirmation. The examinee's ability to falsify and confirm depends on (1) converting the wording of the alternative to that of the text, if required, and (2) comparing propositions between the alternatives and the text. For converting the wording, if the text asserts or denies the alternative verbatim, then its conversion is required. In the current study, encoding conversion was a dichotomous variable to operationalize this effect. In addition, Anderson's (1972) level of text-question transformation was used as an elaboration of encoding conversion.

For comparing propositions, an item would be difficult when the alternatives contain several propositions, each of which must be compared to several propositions in the text. In the current study, propositional comparisons were measured by the natural logarithm of the number of propositional comparisons required to falsify (or confirm) an alternative. For convenience, it was assumed that propositions in the relevant text (see above) are compared exhaustively.

Method

Subjects and Materials

Item response data for the ASVAB and for items that were calibrated for the computerized adaptive test (CAT) version were obtained from a large sample of military applicants for six ASVAB and six CAT forms. All items were administered in booklets. Each applicant received one ASVAB form and one CAT form according to a counterbalanced design for item calibration. Seventy-five paragraph comprehension items were selected from two CAT booklets and the six ASVAB forms. The ASVAB items were selected by two criteria: (1) the full item would fit on the computer screen and (2) only one question was asked per paragraph. A total of 29 ASVAB items met the criteria. The same criteria were applied to the two CAT booklets, resulting in selection of a total of 46 items.

To obtain item parameters for the selected items, data from 12 groups were linked by common items, using the LINLOG program (Whitely & Nieh, 1982) for the LLTM. Estimation of the item parameters in LINLOG is based on Fischer and Formann's (1972) steepest descent algorithm for conditional maximum likelihood. To ensure that the item parameters would not be confounded by item order on the test,

only examinees who reached all items were selected. Thus, the selected sample was probably somewhat above average. However, the ability distribution in the sample does not bias the estimation of item parameters for latent trait models.

Cognitive Variables

The cognitive variables were selected to reflect the measures of processing in the various models. The variables were derived from four sources: (1) propositional analyses of the item texts by two raters; (2) scores on the item response alternatives by two raters; (3) linguistic surface structure variables on the text, question stem, and alternatives, from the STYLE program (Cherry & Vesterman, 1980); and (4) the Kucera-Francis word frequencies for the text, question stem, and alternatives.

For the first data source, two raters independently propositionalized the text of the 75 items. The propositional analysis was based on Kintsch and van Dijk's theory as operationalized by Bovair and Kieras (1981) and Turner and Green (1978). Both raters were graduate students in clinical psychology who were selected due to their high verbal scores on the Graduate Record Examination. Several scores were derived from the propositions, including number of propositions (total), predicate propositions, modifier propositions, connective propositions, and unique arguments. The densities for these five scores were determined by dividing the number of propositions (or arguments) by the number of words in the text.

For the second data source, the raters scored several variables on the alternatives and also propositionalized the alternatives. The variables that were scored included: (1) the number of propositions in the text relevant to determining the correct answer; (2) the truth status of each alternative (confirmed, falsified, or neither); (3) encoding conversion of arguments, a dichotomous variable scored if the arguments of the alternative required inferring synonyms to compare to the text; (4) encoding conversion of predicates, a dichotomous variable scored if the predicates of the alternative required inferring synonyms to compare to the text; (5) number of propositions in each alternative; (6) Anderson's comprehension level (scored 1–6, where 1 = verbatim confirmation of correct answer in the text and 6 = an inference required to confirm the correct answer by the text); (7) Drum et al.'s plausibility variable and (8) Drum et al.'s knowledge variable.

For the third data source, several variables were selected from STYLE as needed in the various models. These included (1) number of words; (2) Flesch's (1948) reading grade level; (3) percent of content words including non-auxiliary verbs, nouns, adverbs, and adjectives; (4) sentence length; and (5) percent of new content words in the question stem and alternatives.

For the fourth data source, the Kucera and Francis (1967) norms were applied to the text, question stem, and alternatives to reflect semantic memory accessibility. Word frequencies were obtained in two ways, for all the words and for just the content words.

Design

Several cognitive models of item difficulty were examined using the LLTM. Program LINLOG for the LLTM not only provides goodness of fit indices for each model, but also calibrates component item parameters that can be used to bank the item by its cognitive characteristics. The parameters also can be used to interpret the effect of each variable on item difficulty, because the model parameters are essentially a mathematical model of item difficulty as scaled by the logistic latent trait model. The parameters of the LINLOG model are analogous to regression weights for the prediction of item difficulty by the cognitive variables. The model fit index, as shown below, has an interpretation similar to the squared multiple correlation.

Results

Descriptive Statistics

The means, standard deviations, and rater reliabilities for the propositional variables are pre-

Table 1
Means, Standard Deviations, and Rater Reliabilities

Variable	Mean	S. D.	Reliability
Propositional Analysis			
Modifier Propositions	11.63	7.47	.94
Predicate Propositions	7.80	4.76	.92
Connective Propositions	10.45	6.17	.93
Total Propositions	29.88	17.07	.98
Arguments	26.01	13.93	.96
Decision Process			
Confirmability	.94	.24	.93
Falsifiability	.63	.48	.73
Encoding Conversion, Correct Answer	.46	.50	.42
Encoding Conversion, Distractors	.29	.45	.26
DCC Plausibility, Correct Answer	.96	.20	.91
DCC Plausibility, Distractors	.20	.40	.97
DCC External Knowledge, Correct Answer	.16	.37	.97
DCC External Knowledge, Distractors	.19	.39	.98
Anderson Level, Correct Answer	4.50	1.21	.52
Anderson Level, Distractors	4.74	1.03	.42

sented in Table 1. Reliabilities were determined by Pearson correlations for continuous variables and by percentage agreement for categorical variables. It can be seen that rater reliabilities of these variables are quite high. Table 1 also presents statistics for the decision process variables that were scored by raters. The reliabilities are also generally high, but with some glaring exceptions. Encoding conversion and the Anderson levels of questioning were not scored reliably, hence the raters were questioned about their scoring criteria. For encoding conversion, one rater applied a much more stringent criterion than was given on the scoring guidelines. Thus, for this variable only the data from one rater were used in the models. For the Anderson variables, the source of unreliability was that the six levels could not be scored as mutually exclusive categories. Two separate variables were then created. The first variable was *paraphrase* level, using categories 1–4 (listed above). The second variable was a dichotomous variable for *reasoning*, scored 1 if either deduction or induction (categories 5–6) was required, and 0 otherwise. The plausibility and external knowledge variables in Table 1 were scored for the Drum et al. (1981) model.

Model Comparisons

Several models for paragraph comprehension items were evaluated by their prediction of item difficulty on a latent trait scale. Each LLTM entails a unique design matrix that consists of the scores for each item on each variable in the model plus a unity vector for the normalization constant. Thus, the design for a three-variable model would be a 75 (item) × 4 (variable) matrix. Embretson (1983a) has proposed several tests to evaluate fit for LLTMs. Shown in Table 2, for example, are the log likelihoods of the data under each model, and the chi-square goodness of fit for each model as compared to a null model. The null model postulates that all items have equal difficulty, hence no reliable information about item differences can be modeled. The design matrix includes only a unit vector for the normalization constant. Also included in the table is an incremental fit index (Embretson, 1983a) which locates the model on a scale from 0 to 1. It gives the amount of information that is predicted by the model, relative to the maximum information that could be predicted by any LLTM (i.e., the difference between the null model and the Rasch model). The index is roughly similar in magnitude to a squared Pearson correlation coefficient.

Table 2
Goodness of Fit for Comparison Models

Models and Variables	Log Likelihood	Chi Square	df	Fit Index
Null	-27647.30			
Rasch	-24007.67	7279.26	74	
Number of Words	-27605.67	83.26	1	.01
Reading Grade Level	-27584.81	124.98	1	.02
Word Frequency	-27544.63	205.34	1	.03
Drum, Calfee & Cook Models				
DCC1- Original:	-26996.23	1302.14	12	.18
DCC2- Expanded:	-26759.00	1776.60	14	.24
Anderson Models				
A1-- Original: Level 1-6	-27234.19	826.11	1	.11
A2-- Truncated: Level 1-4	-27494.05	306.50	1	.04
A3-- Final: Levels 1-4, 5-6	-26991.83	1310.94	2	.18
Mitchell-Like Propositional Model:	-27350.20	594.10	5	.08

Table 2 presents the goodness of fit and the fit index for three single-variable models. Number of words provides significant prediction but accounts for only 1% of the information that could be modeled. Similarly, both the Flesch reading grade level and word frequency for the text provide significant prediction, but account for little information. A logarithmic transformation on the word frequencies offered little change in prediction; therefore, the untransformed frequencies were used throughout the analyses.

Table 3 shows the goodness of fit and the fit index for several models that have been proposed in other studies of paragraph comprehension. Two variants of the Drum et al. model are presented which are not exact replications because some variables were unavailable in the current study. Model DCC1 consisted of the following variables: Paragraph Variables—word frequency, percent content words, actual information; Question Variables—word frequency, percent content words; Correct Alternative—percent new content words, external knowledge, percent content words; Distractors—plausibility, percent new content words, percent content words. Additionally, average word frequency was calculated for the set of alternatives. Model DCC1 contains 12 of 16 variables that were presented by Drum et al. in Table 5 of their study. The current study omits the percent of content-function words in the text and in the question stem, and also omits percentage of new content words for the stem. Last, word frequencies for the correct answer and the distractors were averaged in a single variable. The four predictors that were omitted from DCC1 accounted for an average of 22% of the explained variance in Drum et al. Because the mean squared multiple correlation was .72, deletion of the four predictors should reduce the correlation to approximately .56. In the current data, however, substantially lower prediction was achieved. The fit index reached only .18. Model DCC2 added two variables to DCC1, argument redundancy and word frequency in the text, that were discussed by Drum et al. but were not included in their final models. Substantially better fit was obtained than with DCC1.

The goodness of fit and the fit index for the Anderson models are also presented in Table 3. Model A1 simply uses the numerical designation of the six types of questions presented above as a continuous variable (Anderson levels 1–6). In Model A2, only paraphrase level was used as a predictor (levels 1–4). This model does not fit as well as A1. Model A3 is a two-variable model, using A2 and a second dichotomous variable for reasoning (scored positively if classified as level 5 or 6). Better fit was achieved by A3 than by either A1 or A2. Thus, good prediction was achieved by treating the Anderson model as two variables.

Table 3
Goodness of Fit for Propositional Models of Paragraph Comprehension

Models and Variables	Log Likelihood	Chi Square	df	Fit Index
P1-- Basic Propositional:	-27365.86	562.88	3	.08
P2-- Propositional Type:	-27045.39	1203.82	5	.17
P3-- Propositional Type Revised:	-27156.23	982.14	5	.13
P4-- Final Propositional Type:	-27015.81	1263.98	6	.17

Finally, Table 2 also presents a model like Mitchell's which contains one decision variable (encoding conversion) and four propositional variables (propositional density, argument density, word frequency, and number of words). This model fits better than the single variable models, but prediction was lower than the Drum et al. and Anderson models.

Table 3 shows four text models that were derived from the propositional analysis. The first model, P1, consists of propositional density, argument density, and total word frequency. Significant prediction was achieved, particularly as compared to the model in Table 2 that contained only total word frequency. P1 achieves the same level of prediction as the Mitchell-like model, even though encoding conversion has been deleted. Model P2 consisted of modifier density, predicate density, connective density, argument density, and total word frequency. By separating propositions by type, model P2 achieves a large increase in fit compared to models using only a single combined propositional density measure. An inspection of the weights revealed that the magnitude and direction of the weights varied by propositional type (discussed below).

Model P3 is identical to P2, except that the word frequency index was computed only on the content words, rather than all words. This model is theoretically superior because numerous non-content words (i.e., articles, prepositions, auxiliary verbs, etc.) are not added in with the word frequencies of the arguments and predicates in the text. Interestingly, worse fit was achieved with this model. In P4, the percentage of content words was added to the five variables of P3. The fit was then comparable to P2, which further suggests the importance of surface structure in the total word frequency index.

The newly proposed model for paragraph comprehension outlined earlier was subjected to several intermediate analyses to evaluate the impact of specific measures. In these intermediate analyses, variables were evaluated by the impact on goodness of fit, as compared to a hierarchically nested model that consisted of the text coherence variables. The text coherence processes are postulated to be represented by the Final Propositional Type Model, P4. The impacts of the various postulated decision processes were examined by the increment in χ^2 as compared to the text coherence processes. Adding the lexical encoding processing for the distractors, as measured by word frequency, significantly increased fit ($\Delta\chi^2 = 43.44$, df = 1), but the lexical encoding processing for the correct answer did not significantly increase fit ($\Delta\chi^2 = 1.08$, df = 1). Further, the impact of the text mapping process, as measured by the proportion of text that is relevant to evaluating the alternatives, also significantly increased fit ($\Delta\chi^2 = 224.18$, df = 1).

The evaluation process for the response alternatives was examined separately for the two stages, falsification and confirmation. The falsification process was measured by the probability of guessing correctly among the alternatives that cannot be falsified (i.e., 1 minus the reciprocal of the number of non-falsifiable alternatives plus the correct answer). Adding this variable to the text coherence processes strongly increased fit ($\Delta\chi^2 = 116.52$, df = 1). Then two variables, encoding conversion and propositional comparisons, were examined for possible additive effects by adding each variable as a predictor to the preceding equation. It was

found that propositional comparisons strongly influenced fit ($\Delta\chi^2 = 147.26$, df = 1), while encoding conversion had a small but significant effect ($\Delta\chi^2 = 13.10$, df = 1). Then, the two variables were examined for possible interactive effects with falsification. In these analyses, the falsification probability was multiplied by a fraction that represents the ease of comparing propositions and the ease of encoding conversion. For propositional comparisons, the fraction was the reciprocal, while for encoding conversion, the fraction was the reciprocal of 1 plus the encoding conversion index. A positive finding indicates that the interactive variable occurs during the same stage as falsification. Simple falsification, as reported above, had a strong effect on fit ($\Delta\chi^2 = 116.52$, df = 1). Although the interaction of propositional comparisons with falsification did not affect fit as much as falsification alone ($\Delta\chi^2 = 47.66$), the interaction of encoding conversion with falsification had a stronger effect ($\Delta\chi^2 = 161.24$, df = 1). In summary, these results suggest that encoding conversion has an interactive effect with falsification, but that propositional comparisons has an additive effect.

The confirmation process was measured by a dichotomous variable that reflected whether the correct answer was actually confirmed by the text. When added to the text coherence process model, confirmation strongly increased fit ($\Delta\chi^2 = 585.62$, df = 1). When added to the comparison model of text coherence processes and confirmation, both propositional comparisons ($\Delta\chi^2 = 142.68$, df = 1) and encoding conversion ($\Delta\chi^2 = 128.88$, df = 1) had significant effects. Next, encoding conversion and propositional comparisons were examined for possible multiplicative effects. The interaction of confirmation and encoding conversion provided an even stronger increment in fit than did the simple confirmation variable ($\Delta\chi^2 = 729.60$, df = 1). However, the interaction of confirmation and propositional comparisons did not increase fit over the text coherence processes as much as the simple confirmation variable ($\Delta\chi^2 = 446.04$, df = 1). Thus, as for falsification, the strongest effect for encoding conversion or confirmation was interactive, while for propositional comparisons the strongest effect was additive.

Table 4 presents the goodness of fit and the fit index for the combinations of the various decision process variables described above (Tables 5 and 6 give a complete listing of these variables). It can be seen that D1, the Complete Decision Model, has a fit index of .25. In this model, falsification and confirmation are multiplied by encoding conversion, as supported by the preceding results. For comparison, the Complete Text Model T1 (also presented as P4 above), shown in Table 4, has a fit index of .17. Thus, the decision process variables provide better fit than the text process variables. Also shown in Table 4 is the goodness of fit for the model F1, which combines T1 and D1. F1 achieves a fit index of .37. In this model, propositional comparisons between the text and alternatives were positively related to item difficulty, indicating that

Table 4
Goodness of Fit for Proposed Models of Text and Decision

Models and Variables	Log Likelihood	Chi Square	df	Fit Index
T1-- Proposed Text Model:	-27015.81	1263.98	6	.17
D1-- Proposed Decision Model:	-26727.87	1838.86	7	.25
D2-- Proposed Decision Model, Substituting Anderson Variables:	-26618.67	2057.26	7	.28
F1-- Proposed Full Model T1 + D1 Variables:	-26317.07	2660.46	13	.37
F2-- Proposed Full Model--Anderson Variables T1 + D2 Variables:	-26305.22	2684.16	13	.37

many comparisons are characteristic of more difficult items. This model differs from the next model to be presented only in exchanging the two propositional comparison variables for the alternatives.

A second decision model D2 is shown in Table 4 in which the Anderson reasoning variables replace the D1 propositional comparison variables for both the correct and distractor alternatives. The reasoning variables simply scored the presence or absence of Anderson's induction-deduction levels (5–6). Further, the paraphrase level variable replaces encoding conversion as the multiplier of falsification and confirmation. Model F2 combines the Complete Text and Decision Models T1 and D2, achieving a fix index of .37.

LLTM Calibrations

Table 5 shows the LLTM calibrations for the final Complete Model F2 that combines the text and decision models T1 and D2. The η weights, the asymptotic standard errors, and the associated t values are presented. The η weights should be interpreted in a manner similar to that of unstandardized regression coefficients. That is, the weights reflect the impact of the variable, controlling for the effects of the other (possibly correlated) variables in the model. The rs in the table are not from the LLTM. They are zero-order correlations of Rasch item difficulty with the variables, allowing the detection of direct versus suppressor effects.

For the Text Model T1, the variables that contribute significantly to predicting item difficulty are modifier and predicate density, percent content words, and (marginally) argument density. The direction of the impact is worth noting, as well as where the r and η differ in sign. Modifier propositional density and percent content words have a significant positive relationship to item difficulty for both the model and the zero-order relation. Thus, items in which the text has relatively many modifier propositions and a high proportion of words with semantic information tend to be more difficult items. However, the opposite significant impact is found for connective propositional density, where items in which the text has very few connective propositions are more difficult. There are negative, but non-significant, weights for argument density and predicate density. However, argument density increases with item difficulty in the zero-order relation, reflecting a suppressor role in the model. Text content word frequency shows only a very small and non-significant positive relation.

For the Decision Model D2, nearly all of the

Table 5
Complexity Factor Weights for Proposed Model F2

Variable	r	η	SEη	t
Text Model (T1)				
Modifier Propositional Density	.174	2.30	.58	3.91 **
Predicate Propositional Density	-.020	-.33	.56	-0.59
Connective Propositional Density	-.205	-3.88	.53	-7.34 **
Argument Density	.161	-.88	.48	-1.82
Text Content Word Frequency	.014	.07	.11	0.69
Percent Content Words	.272	.54	.27	1.97 *
Decision Model (D2)				
Percent Relevant Text	.175	.20	.02	8.91 **
Falsification	-.186	-1.51	.70	-2.15 *
Confirmation	-.405	-2.72	.41	-6.59 **
Word Frequency, Distractors	-.274	-.43	.16	-2.71 **
Word Frequency, Correct	-.121	.27	.15	1.82
Reasoning - Distractors	.112	-.29	.17	-1.75
Reasoning - Correct	.356	.55	.18	3.15 **

* ($p < .05$)
** ($p < .01$)

variables contribute significantly to predicting item difficulty. It can be seen that text mapping, as measured by the percent of text relevant to the alternatives, is positively related to item difficulty. Thus, the more relevant the text, the more difficult the item. The newly created falsification and confirmation variables (i.e., the previous ones multiplied by the reciprocal of paraphrase level) are negatively related to item difficulty. That is, the lower the likelihood that the distractors are falsified and the correct answer confirmed, the greater the difficulty of the items.

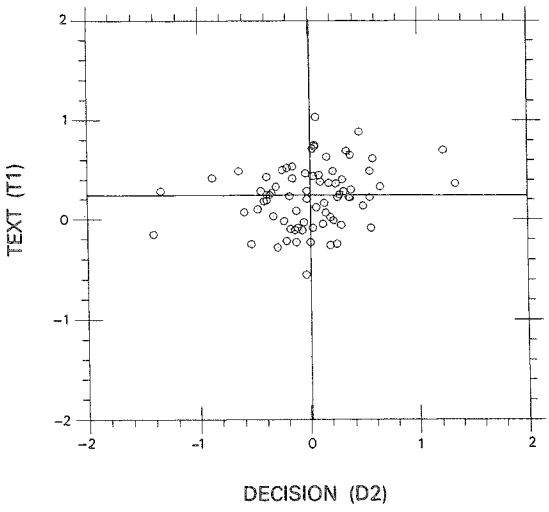
Interestingly, two sets of variables in Model D2 showed suppressor effects, where two variables were correlated with difficulty, but the one with the weaker correlation yields the opposite sign in its contribution to the model. Word frequency had opposing effects on the correct answer and the distractors in D2, even though both had a negative zero-order relation. If the distractors consist of infrequent words, the item is more difficult. If the correct answer consists of infrequent words, the item is easier in the model (or more difficult for frequent words), but this effect is less pronounced. By contrast, the reasoning variables had opposing effects on the correct answer and distractors in D2, but both were positively related to difficulty by the zero-order correlations. In D2, a greater reasoning score involving induction-deduction was positively related to difficulty for the correct answer, but carried a negative weight for the distractors.

Construct Validity

The cognitive model that was developed above explicates the construct representation of the item bank by specifying the sources of cognitive complexity that are related to item difficulty. However, a more precise examination of the cognitive characteristics of the item bank may be obtained by calibrating complexity scores for individual items and examining the descriptive statistics.

Figure 3 shows the relationship between difficulty of text representation (T1) and difficulty of decision (D2). It can be seen that text representation and decision complexity are relatively independent sources of variation in the item bank ($r = .23$). A similar relation was obtained for models T1 and D1 ($r = .18$). Thus, item writers have manipulated these qualities relatively independently. Figure 3, however, also shows that items with extreme configurations of text representational complexity and decision complexity exist in the item bank. The two axes drawn within the figure represent the medians of the two predicted processes. It is clear that items with a specified cognitive characteristic could be obtained from any of the four quadrants. For example, items could be selected to cause a difficult decision, while leaving text difficulty to vary, by selecting only items to the right of the vertical line. A specification such as this could be used to guide item writing by calibrating the difficulty of an item on the various process variables prior to empirical tryout by using the component latent trait parameters.

Figure 3
Scatterplot of Item Difficulties Predicted by Text Representation (T1) and Decision (D2) Processes



Because the items were rather heterogeneous for a source of cognitive complexity, the unidimensionality of the extreme items was tested. Items that were difficult with respect to the text but not the decision formed one subset, while items that were difficult with respect to the decision but not the text formed the second subset. The Martin-Lof

(1974) test for unidimensionality yielded a highly significant χ^2 of 2799.28 with 357 degrees of freedom. Thus, the two item sets measure different abilities.

Table 6 compares the cognitive characteristics of the ASVAB items to the newer CAT items. A multivariate analysis of variance was conducted, with the scores on the cognitive variables as the dependent variables and the item source (ASVAB vs. CAT) as the grouping variable. Table 6 presents the results for the predictors in T1, D1, and D2. It can be seen that ASVAB and CAT items are not significantly different on the text representation process ($p = .130$).

However, somewhat different results were obtained from the decision component comparisons. It can be seen that the sources of cognitive complexity in model D1 differed significantly ($p = .047$) between ASVAB and CAT. The univariate analyses of variance show that to falsify the distractors, CAT items require significantly more comparisons between propositions in the alternatives and those in the text. Two marginally significant differences were also found. The CAT items require more comparisons to confirm the correct answer and the percentage of relevant text for the CAT items was greater.

The second set of decision variables, D2, showed no significant differences. In this set, the propositional comparison variables had been replaced by the reasoning variables.

Discussion

Paragraph comprehension items appear in tests

Table 6
Multivariate and Univariate Analysis of Variance for ASVAB and CAT on Text and Decision Models

Variable	Means		F Ratio	P Value
	ASVAB	CAT		
Text (T1)				
Multivariate	-----	-----	1.720	.130
Modifier Density	.18	.19	.468	.496
Predicate Density	.12	.13	.492	.485
Connective Density	.17	.17	.260	.612
Argument Density	.42	.42	.000	.986
Content Word Frequency, Text	1050.01	841.42	.983	.325
Percent Content Words	.55	.59	8.850	.004
Decision (D1)				
Multivariate	-----	-----	2.180	.047
Percent Relevant Text	.58	.68	2.831	.097
Falsification	.28	.23	2.344	.130
Confirmation	.63	.57	2.055	.156
Word Frequency, Distractors	921.02	1025.92	.482	.490
Word Frequency, Correct Answer	769.07	819.84	.096	.757
Propositional Comparisons, Dist.	4.97	5.45	7.562	.008
Propositional Comparisons, Corr.	3.96	4.33	3.802	.055
Decision (D2)				
Multivariate	-----	-----	.810	.576
Percent Relevant Text	.58	.68	2.831	.097
Falsification	.05	.04	1.710	.195
Confirmation	.23	.20	1.989	.163
Word Frequency, Distractors	921.02	1025.92	.482	.490
Word Frequency, Correct Answer	769.07	819.84	.096	.757
Reasoning, Distractors	.78	.79	.006	.937
Reasoning, Correct Answer	.55	.67	1.121	.293

with rather different measurement goals. That is, they appear both in measures of verbal reasoning and of reading achievement. This suggests that the items can be varied to measure rather different constructs. However, the cognitive characteristics of paragraph comprehension items, which could be varied according to the measurement goals, have not been studied sufficiently to allow systematic variation of item content in accordance with a specified measurement goal.

The current study developed a processing model to quantify the sources of cognitive complexity in multiple-choice paragraph comprehension items. The results indicate that relatively good prediction could be achieved from the proposed processing model which included stages for Text Representation and Decision. The best model consisted of six text representation variables and seven decision processing variables, yielding a fit index of .37, which is comparable to a multiple correlation of .61. Although the magnitude of the fit index indicates that item difficulty is not fully predicted by the model, the index is high enough to indicate that substantial prediction has been achieved.

Perhaps the most significant finding from the model is that the decision process influences item difficulty substantially more than the text representation process. The decision variables alone achieved a fit of .28 while the text representation variables achieved a fit of only .17. Thus, item difficulty of the paragraph comprehension test depends more on the response decision than the paragraph. This finding may have important implications for the predictive validity of paragraph comprehension tests, particularly if the text representation process and decision process are associated with different abilities. It would appear that two recognized factors of ability have been identified: verbal ability (or more specifically, language comprehension and lexical knowledge) and reasoning ability. This will be considered more fully below.

Interpretations given below about the influence of specific variables on item difficulty must be tempered by noting that the weights resemble unstandardized regression coefficients. Thus, the magnitudes of the weights are dependent on the scale of measurement for each variable and cannot be compared across variables. Further, the significance of variables is influenced by intercorrelations among the predictors, just as in multiple regression. A variable that is highly correlated with other variables in the model may fail to have a significant weight, even though it may in fact be a factor that could be manipulated to change item difficulty. Alternatively, a variable that has a significant weight may not necessarily be a factor that determines item difficulty if it is correlated with an unmeasured variable that does determine item difficulty.

Specific Text Influences

A major finding from having separated propositions into three types was that the prediction of item difficulty increased (Table 3) and that the effect of propositional density depended on the type of proposition (Table 5). Connective density had a consistent and high negative weight in the propositional analysis models and by the zero-order correlation; that is, greater density was found with easier items. The direction of this effect is surprising because higher propositional content per total paragraph text might be expected to make comprehension more difficult. However, the expectation is probably misleading. Connective propositions relate or link other facts or propositions together (e.g., their arguments are often other propositions) and are thought to be important to providing text cohesion (Turner & Green, 1978). For example, connectives indicate relationships between propositions of cause, purpose, conjunction, disjunction, condition, concession, contrast, or circumstance. Additionally, because important propositions are better remembered, it has been thought that they are processed more frequently (Kintsch & Keenan, 1973; Kintsch et al., 1975). This "levels" effect generally refers to superordinates containing information relevant to the entire text, which are processed more frequently because arguments overlap between successive processing cycles (Miller & Kintsch, 1980). The superordinates that usually link text for coherence are connectives and, to a lesser extent, predicates. However, if few connectives are given or some are omitted, the reader may

have to supply them in order to comprehend. This suggests that texts with few connective propositions may require additional processing resources or more inferences. It is nearly axiomatic in comprehension research that text becomes more difficult as the reader must supply more inferences to bridge gaps and make the text coherent. Because the negative relation to item difficulty was relatively large, this result warrants further study.

Predicate propositions did not have a significant relation in the model and, further, the zero-order correlation with item difficulty was near zero. Many predicates in the list of propositions do not seem to function as superordinates, and thus they may not participate as frequently in processing cycles as do connectives. Still, why dense terms of action or being are not more highly related to difficulty in either direction is not clear. Until a more detailed examination of predicates is made, it might be concluded that predicate density does not influence item difficulty.

In contrast, the density of modifiers was positively related to item difficulty in the final model. Further, the occasional differences in weights for modifier propositional density probably result from high correlations with other propositional variables such as argument density, which is also positively related to item difficulty. Since modifiers are not usually the superordinates that contribute to coherence, the results suggest that a major source of difficulty for these items is the processing of dense modifiers that qualify, quantify, or indicate a quality of an argument.

The findings for argument density and percent content words were also consistent with the idea that an increased demand on the reader's processing resources is positively related to item difficulty. The introduction of new material or new word concepts should also place greater demands upon the reader's resourcefulness. The density of unique arguments can also be related to the introduction of new material and was positively related to difficulty by its zero-order correlation. However, it yielded a negative weight in the final model because of its supressor effect of being less highly correlated with difficulty than the other text variables. In the case of the percent content words measure, the introduction of new material refers to the density of meaning-bearing words. The relation to item difficulty was positive in the models and by the zero-order correlation. The density of words carrying meaning is relatively simply reflected by this content words measure because it merely strips away function words, such as articles and prepositions.

Finally, word frequencies should also reflect the introduction of new word concepts as described above. For paragraphs, word frequency was only slightly related to difficulty, even though the sign was positive. As with Drum et al. (1981), predicting difficulty from word frequency was less important for the paragraph than for the response alternatives or question stem. The poor relationship with paragraph difficulty probably means that this item format allowed sufficient exposure time so that it reflected words already encoded, rather than initial encoding. More local measures, such as reading time, would show more significant effects because they better tap initial encoding; once encoded the effect should be less, as in the present work (see Kieras & Just, 1984).

Specific Decision Influences

All three decision events had significant impact on item difficulty. The encoding processing stage was measured by word frequency for the correct answer and word frequency for the distractors. While both had a negative relation by zero-order correlations, these variables had opposing effects on item difficulty in the models. That is, more difficult items tended to be ones with infrequent words in the distractors, but more frequent words for the correct alternative. These findings reflect the inconsistency mentioned above, where the effect of a variable changes in different models. These findings probably result from a kind of contrast effect. For example, studies have found that items are easier if the correct answer is more complex than the distractors (e.g., Chase, 1964).

These results indicate a global effect for text linking-mapping as a separate evaluation event in the models. The percent of relevant text between the paragraph and alternatives was positively related to difficulty and significantly improved good-

ness of fit. An additional positively related measure of text linking-mapping used in Model D1 was the number of propositional comparisons between the alternatives and the text. The Anderson reasoning variables reflected the greater difficulty associated with the presence of induction-deduction between the text and the alternatives. The more highly correlated reasoning score for the correct answer bore a positive weight, and that for the distractors carried a suppressor effect. All of these variables reflect the complexity or volume of information required to evaluate alternatives against the text and demand more processing resources of the examinee.

The information processing model was expanded by adding a two-stage falsification-confirmation response evaluation process. Several models of the falsification stage were compared. It was found that falsification success, when measured as the probability of guessing correctly from the remaining (non-falsifiable) alternatives, significantly predicted item difficulty. The Anderson paraphrase level variable interacted with falsification, while text mapping, as measured by number of propositional comparisons, had additive effects. Thus, falsifying an alternative depends on transforming the alternatives to the text, while the number of comparisons has a more global effect on item difficulty. A large number of comparisons probably places an overall demand on working memory, but not on item differences in falsifiability.

In the final models, the confirmation and falsification stage variables were transformed by multiplying them by the reciprocal of the paraphrase level. The confirmation stage was measured by classifying items according to whether the text confirmed the correct answer. Confirmation had a more substantial effect on item difficulty than did falsification. Apparently, ambiguity in confirming a correct answer creates difficult items. As for falsification, encoding conversion had an interactive effect on confirmability, while number of comparisons had an additive effect.

Model Comparisons

The propositional analysis models (P1–P4) performed better than two traditional indices of text difficulty, reading grade level and word frequency. These models were clearly superior to a popular measure of readability alone, the Flesch (1948) index, which accounted for only 2% of the information to be modeled. In contrast, the propositional analysis models accounted for 8% to 17% of the information. Thus, as anticipated from Kintsch and van Dijk's approach, item difficulty is influenced by the propositional structure of the text, particularly by connective and modifier propositions. A five-variable Mitchell-like model fit as well as the three-variable basic propositional model (P1) containing propositional density, argument density, and word frequency, even though the Mitchell-like model included encoding conversion. Some additional propositional models (P2–P4) yielded even better fit than these propositional models in having separated total propositions into predicates, modifiers, and argument densities, as well as incorporating percent content words.

Comparisons to the Anderson and Drum et al. models generally supported the relative importance of the decision process over the text representation process. The Anderson model measured the extent to which a transformation was required between the text of the paragraph and the alternatives; that is, it reflects decision text mapping and evaluation stages. While the Anderson A3 model contained only two predictors, it fit as well as much larger models containing text representation variables. Although Drum et al. had provided good prediction for paragraph comprehension items on reading achievement tests, their models did not well represent the decision process. The present approximation to their model did not fit as well as the final model incorporating both text and decision processes with nearly the same number of variables.

Implications For Measurement

The cognitive characteristics of test items have an intrinsic role in the construct validity of a test. As has been shown elsewhere (Embretson, 1983b), construct validity may be conceptualized into two separate aspects, construct representation and nomothetic span. Construct representation research ex-

amines the construct involved in item solving from identifying the components, strategies, and knowledge structure that underlie performance. The cognitive characteristics of items directly influence construct representation by controlling the item stimulus factors that influence cognitive processing. Nomothetic span concerns the utility of an index of item solving (i.e., the test score) for measuring individual differences by examining its relationships to other tests, criterion measures, and demographic variables. Cognitive characteristics also influence nomothetic span, if the various components, strategies, and knowledge structures define somewhat different aspects of individual differences (see Embretson, Schneider, & Roth, 1986).

One useful outcome of the present study for measurement is an explication of the cognitive design principles in the item bank. Although no specifications for the cognitive characteristics that were measured here currently exist for paragraph comprehension items, some definite design principles emerged from the data. First, as indicated above, the decision component had a larger impact on item difficulty than the text component. This implies that the manipulation of propositional density and type, as well as word familiarity and percentage of content words, did not have much effect in the item bank. Second, the complexity of the text representational component was relatively independent of the complexity of the decision component. LLTM calibrations for items on text and decision had small correlations (r = .18 and .23). These results show that test developers have constructed the items so that they may be independently influenced by the text component and the decision component.

These results imply that items could be explicitly constructed to be difficult on either or both components. The specific weights for the variables of text representational complexity and decision complexity suggest how to manipulate item content to control the source of item complexity. That is, the weights given for the cognitive model developed above can be used to estimate the complexity of the item on cognitive components.

Similarly, the results suggest how items may be selected to control cognitive complexity. The paragraph comprehension subtest from the ASVAB is particularly amenable to improvement (see Moreno, Wetzel, McBride, & Weiss, 1984). The lower reliability of the paragraph comprehension subtest compared to the other ASVAB subtests stands in contrast to the comparable figure of .8 to .9 for the well known Gates-MacGinitie (1978) reading test.

However, the data also indicated that nomothetic span may be influenced by the source of cognitive complexity. It was found that items that were primarily complex on text measured a different ability than items that were complex primarily on decision. Thus, the trait that is measured by a specific set of items will depend on the balance between text and decision as sources of difficulty.

The heterogeneity of the item bank with respect to sources of cognitive complexity, coupled with different abilities, could lead to difficulties for the implementation of computerized adaptive testing for paragraph comprehension tests. Because the items received by examinees vary in content under adaptive testing, it is possible that examinees could receive item sets that are extreme on text or decision, thus leading to problems in score equating. Further research is clearly needed to determine the magnitude of this problem, as the current results are only suggestive. If, in fact, the nomothetic span is appreciably altered by sources of cognitive complexity in the items, additional balancing of cognitive characteristics between adaptive tests or more precise item specifications may be needed.

A practical implication of the procedures developed in this work is the potential for creating a computer program similar to STYLE (Cherry & Vesterman, 1980) which would allow test item writers to "try out" the text of a new item on-line. Such a program could return a predicted difficulty from automated text scorings based on established predictor weights. If the item characteristics did not fit some specification, then the test item could be modified immediately and a new on-line analysis could be tried until the specification was met. This exercise would occur prior to actually administering it to a group of examinees, thereby saving the expense of a test administration and item calibration based on a large group of examinees. Of course, text processing technology and research on com-

ponent models for paragraph comprehension items is needed before this potential application can be realized.

References

Anderson, R. C. (1972). How to construct achievement tests to assess comprehension. *Review of Educational Research, 42*, 145–170.

Bovair, S., & Kieras, D. E. (1981). *A guide to propositional analysis for research on technical prose* (Technical Report No. 8). Tucson AZ: University of Arizona.

Chase, C. L. (1964). Relative length of options and response set in multiple choice items. *Educational and Psychological Measurement, 24*, 861–866.

Cherry, L. L., & Vesterman, W. (1980). *Writing tools—The STYLE and DICTION programs* (Computing Science Technical Report No. 91). Murray Hill NJ: Bell Laboratories.

Drum, P. A., Calfee, R. C., & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly, 16*, 486–514.

Embretson, S. E. (1983a, June). *An incremental fit index for the linear logistic latent trait model*. Paper presented at the annual meeting of the Psychometric Society, Los Angeles.

Embretson, S. E. (1983b). Construct validity: Construct representation and nomothetic span. *Psychological Bulletin, 93*, 179–197.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49*, 175–186.

Embretson, S. E. (1985). *Test design: Developments in psychology and psychometrics*. New York: Academic Press.

Embretson, S. E., Schneider, L. M., & Roth, D. L. (1986). Multiple processing strategies and the construct validity of verbal reasoning tests. *Journal of Educational Measurement, 23*, 13–32.

Embretson, S., & Wetzel, D. (1984, April). *Latent trait models for the cognitive components of paragraph comprehension items*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359–374.

Fischer, G. H., & Formann, A. K. (1972, October). *An algorithm and a FORTRAN program for estimating the item parameters of the linear logistic test model* (Research Bulletin No. 11). Vienna: Psychologisches Institut, Universität Wien.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*, 221–233.

Gates, A., & MacGinitie, W. (1978). *Gates-MacGinitie Reading Tests, Survey D, Teacher's Manual*. Boston MA: Houghton Mifflin.

Kieras, D. E., & Just, M. A. (1984). *New methods in reading comprehension research*. Hillsdale NJ: Erlbaum.

Kintsch, W., & Keenan, J. (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology, 5*, 257–274.

Kintsch, W., Kozminsky, E., Streby, W. J., McKoon, G., & Keenan, J. M. (1975). Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior, 14*, 196–214.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*, 363–394.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence RI: Brown University Press.

Martin-Lof, T. (1974). The notion of redundancy and its use as a quantitative measure of discrepancy between a statistical hypothesis and a set of observational data. *Scandinavian Journal of Statistics, 1*, 3–18.

Miller, J. R., & Kintsch, W. (1980). Readability and recall of short prose passages: A theoretical analysis. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 335–354.

Mitchell, K. (1983). *Cognitive processing determinants of item difficulty on the verbal subtests of the Armed Services Vocational Aptitude Battery and their relation to success in Army training*. Unpublished doctoral dissertation, Cornell University.

Moreno, K. E., Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests. *Applied Psychological Measurement, 8*, 155–163.

Pellegrino, J. W., & Glaser, R. (1979). Cognitive correlates and components in the analysis of individual differences. *Intelligence, 3*, 187–214.

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. New York: Academic Press.

Turner, A., & Green, E. (1978). Construction and use of a propositional text base. *JSAS Catalog of Selected Documents in Psychology, 3*, 58. (Ms. No. 1713)

Whitely, S. E., & Barnes, G. M. (1979). The implications of processing event sequences for theories of analogical reasoning. *Memory and Cognition, 7*, 232–331.

Whitely, S. E., & Nieh, K. (1982). *LINLOG* [Computer program]. Unpublished manuscript, University of Kansas, Lawrence KS.

Whitely, S. E., & Schneider, L. M. (1981). Information

structure on geometric analogies: A test theory approach. *Applied Psychological Measurement, 5*, 383--397.

Acknowledgments

Susan E. Embretson has also published under the name Susan E. Whitely. This work was supported by the Navy Personnel Research and Development Center through the Army Research Office and Battelle Memorial Institute Contract No. 0855. The opinions expressed in this article are those of the authors, are not official, and do not reflect the views of the Departments of the Navy or Army.

Author's Address

Send requests for reprints or further information to Susan Embretson, Department of Psychology, University of Kansas, Lawrence KS 66045, U.S.A.