DOCUMENT RESUME

ED 292 875 TM 011 312

AUTHOR Bejar, Isaac I.; And Others
TITLE Cognitive Psychology and the SAT: A Review of Some Implications. Research Report 87-28.
INSTITUTION Educational Testing Service, Princeton, N.J.
PUB DATE Jul 87
NOTE 78p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Cognitive Psychology; *College Entrance Examinations; Diagnostic Tests; Difficulty Level; Higher Education; *Item Analysis; Outcomes of Education; Psychometrics; *Standardized Tests; Test Construction; *Test Validity
IDENTIFIERS *Scholastic Aptitude Test

ABSTRACT

A review of the implications for the validity of the Scholastic Aptitude Test (SAT) of scientific developments accompanying the revival of cognitive psychology provides insights into the importance of such changes. A distinction can be made between a process-oriented or diagnostic test and an outcomes-oriented test such as the SAT. Since the SAT does not aim to be a diagnostic test, the implications of cognitive theories for test construction that would otherwise be applicable are not emphasized. Nevertheless, an analysis based on quantitative items illustrates how tests with a diagnostic orientation could be based on cognitive principles. An explication was developed for the linkage of cognitive principles with psychometric considerations for an outcomes-oriented test. The essence of that linkage is an accounting of item difficulty. The understanding of differential difficulty among items can be viewed as an additional requirement for construct validity, which until recently, focused almost exclusively on an accounting of the co-variation of test scores in terms of "abilities." Although work from this enlarged validational perspective on SAT items is limited, there is work based on similar items that suggests that the SAT will fare well when relevant studies are conducted. Future studies should be based on an enlarged validational framework. A nine-page list of references is included. (TJH)

RR-87-28

RESEARCH REPORT

COGNITIVE PSYCHOLOGY AND THE SAT: A REVIEW OF SOME IMPLICATIONS

Isaac I. Bejar
Susan Embretson
Richard E. Mayer

Educational Testing Service
Princeton, New Jersey
July 1987

Cognitive Psychology and the SAT: A Review of Some Implications

Isaac I. Bejar

Educational Testing Service

Princeton, NJ

Susan Embretson

University of Kansas

Lawrence, KS

Richard E. Mayer

University of California

Santa Barbara, CA

Abstract

That the SAT has not changed fundamentally since its inception without diminishing its pragmatic utility suggests that the SAT is appropriate for its intended purpose. Nevertheless, the social and scientific milieu has certainly changed since the 1920s. This paper reviews the implications of scientific developments accompanying the revival of cognitive psychology. We distinguished early in the paper between a process-oriented or diagnostic test and an outcomes-oriented test such as the SAT. Since the SAT does not aim to be a diagnostic test the implications of cognitive theories for test construction that would otherwise be applicable were not emphasized. Nevertheless, an analysis was presented based on quantitative items that illustrates how tests with a diagnostic orientation could be based on cognitive principles. Most of the paper, however, is devoted to an explication of the linkage of cognitive principles with psychometric considerations for an outcomes-oriented test like the SAT. The essence of that linkage is an accounting of item difficulty, that is, an understanding of the differences in difficulty among items. This understanding, can be viewed as an additional requirement for construct validity, which until recently focused almost exclusively on an accounting of the covariation of test scores in terms of "abilities." In other words, explaining covariation *and* item difficulty are now equally important aspects of validation. Although work from this enlarged validational perspective on SAT items is limited, there is work based on similar items, especially analogy and reading comprehension, which suggests that the SAT will fare well when relevant studies are conducted. Despite this positive outlook additional studies need to be conducted based on an enlarged validational framework. Even though the predictive power of the test is not likely to increase significantly because of this research, it is nevertheless essential as a means of realizing the programmatic implications discussed in the paper.

Executive Summary

That the SAT has not changed fundamentally since its inception without diminishing its pragmatic utility suggests that the SAT is appropriate for its intended purpose. Nevertheless, the social and scientific milieu have certainly changed since the 1920s. This paper reviews the implications for the SAT of the scientific developments accompanying the revival of cognitive psychology. We distinguished early in the paper between a process-oriented, or diagnostic test, and an outcomes-oriented test, such as the SAT. Since the SAT does not aim to be a diagnostic test the implications of cognitive theories for test construction that would otherwise be applicable were not emphasized. Nevertheless, an analysis was presented for quantitative items that illustrates how tests with a diagnostic orientation could be based on cognitive principles. Most of the paper, however, was devoted to an explication of the linkage of cognitive principles with psychometric considerations for an outcomes-oriented test like the SAT. The essence of that linkage is an accounting of item difficulty, that is, an understanding of the differences in difficulty among items. This understanding can be viewed as an additional requirement for construct validity, which until recently focused almost exclusively on an accounting of the covariation of test scores in terms of "abilities." In other words, explaining covariation and item difficulty are now equally important aspects of validation. Although work from this enlarged validational perspective on SAT items is limited, we reviewed work based on similar items, especially analogy and reading comprehension, which suggests that the SAT will fare well when relevant studies are conducted. Despite this positive outlook additional studies need to be conducted based on this enlarged validational framework. Even though the predictive power of the test is not likely to increase significantly because of this research, it is nevertheless essential as a means of realizing some of the programmatic implications discussed in the paper. These programmatic implications are discussed next.

Content Implications

If any changes were to be made to the SAT on account of cognitive psychology they are not likely to involve *removing* the item types that are currently used because performance on those items is widely acknowledged to be a phenomenon well within the scope of interest of cognitive psychology. Moreover, there appears to be no obvious way in which cognitive psychology research can directly improve the measurement of developed verbal ability. For example, Hunt & Pellegrino (1984) have concluded with respect to the assessment of intelligence that the joint application of cognitive psychology and computers was not likely to "extend the range of evaluation if only because the simple vocabulary test is such a good predictor."

The lack of suggestions from cognitive psychology on how to modify the content of tests is probably not a shortcoming of cognitive psychologists, as Sternberg seems to suggest, but rather a poignant indicator that psychometrics and cognitive psychology are concerned with the same phenomena. As noted by Sternberg himself (1984): ...."I would argue that cognitive psychology has provided a valuable *complementary* way of investigating pretty much the same construct psychometricians have been studying all along" (p. 49, italics added).

Granted that both cognitive and psychometric approaches to individual differences are concerned with the same phenomena, their characterization of the students' knowledge and skills is not, as we have pointed out throughout the paper. Because psychometrics often is concerned with ordering students on a continuum, it stresses "how much the student knows." Cognitive psychology, however, does not have similar pragmatic objectives, and instead focuses on a more psychologically motivated description of "what the student knows." This description is governed by theoretical considerations rather than pragmatic ones and may not be directly usable for a psychometric characterization of individual differences, at least with respect to standard psychometric models. Psychometric models such as Item Response Theory (e.g., Lord, 1980), which is now used for equating the SAT, have put psychometric inference on a solid statistical foundation but have required that certain constraints be met that, at least at first sight, do not have psychological import. The challenge then, as we see it, is to blend psychometric and cognitive concerns harmoniously by taking advantage of the framework provided by psychometric models. We would thus insure valid conclusions and, by examining the cognitive literature, we would focus on the processes and representations used by examinees to solve test items.

Implications for Test Development and Test Administration

Just as attention to cognitive psychology has suggested a major reformulation of the concept of validity it may also significantly affect the test development process of existing tests, and in turn may lead to more valid and efficient measurement methods. The impact on test development construction is twofold. On one hand the psychometric response models that are used to explain the data may themselves be partly motivated by some psychological theory of the response process. Second, the process of creating items may itself be enhanced by input from cognitive psychology.

That is, instead of creating and discarding items through an exclusively empirical item it may be possible to design items with specific psychometric characteristics.

Equatability

An important objective of the test developer is to insure that forms are comparable from year to year. This is accomplished by following closely a set of guidelines that specify, among other things, the range and distribution of item difficulty and discrimination. Independent of these statistical requirements the test developer must also follow strict content guidelines. An outcome of linking cognitive task analysis and psychometric modeling is that the test developer has a richer set of item attributes to work with. To the extent that these attributes are based on a valid psychological model of the item solution process, there is a chance of improving comparability of forms from year to year. In the absence of psychological item attributes the test developer relies on categorizations of items whose psychological foundation is not well documented. For example, analogy items are classified along a number of dimensions, including one that sorts the items into a concrete-abstract-mixed trichotomy. Although it is a sensible categorization, its origin is not known.

Predicting Item Difficulty

In addition to improving comparability, a cognitively oriented approach has the potential of helping test developers anticipate the difficulty of items. This would be of significant practical importance. Bejar (1983), for example, has reported that at least with respect to TSWE items there is considerable room for improvement in the prediction of item difficulty. On the other hand, the combination of the test developer's expertise with information about the cognitive demands of the items is a much better predictor of item difficulty than either source of information alone (Bejar, Stabler, & Camp, 1987). Similarly, there is evidence from an ongoing GRE analogy study (Bejar & Enright, in preparation) that cognitive information and test developer expertise are much better predictors of item difficulty than either source alone.

The practical applications of being able to anticipate difficulty include the possibility of reducing the need for pretesting of items. This could come about in at least two ways. First, if estimates of difficulty for unpretested items are available it may be possible to more effectively choose which one to actually pretest on the basis of the gaps that may exist in the item pool. For example, if there are plenty of easy items, then obviously there is no point in pretesting items that are estimated to be easy. Second, if difficulty can be estimated successfully then clearly we have a handle on what makes the items harder or easier to begin with. That knowledge could be put to use while the item is being developed rather than after the fact. This is becoming more feasible as the entire test development process is more and more assisted by computers. To give a simple example, while composing an item the test developer could instantaneously look up the vocabulary load of the item as currently drafted, and revise it if necessary.

Test Administration

Many of the criticisms of the SAT seem to stem from its being a difficult test. Because computerized test administration often relies on item response theory (Lord, 1980) it is feasible through computer administration to adjust the difficulty of the test for each examinee. That is, instead of measuring every examinee with equal precision, which is the usual objective of adaptive tests, a computerized test could be designed, to administer, not in effect to be, the most psychometrically efficient test, but rather a test that would be easier than under normal adaptive procedures. To be sure, this would entail loss of precision of measurement, but the loss could in turn be compensated by lengthening the test.

An application of adaptive testing technology to make the test easier could proceed on its own, independent of any cognitive considerations. Of course, by combining adaptive testing technology with cognitive item analysis we can not only control the real and perceived difficulty of the test but also help to maintain the the validity of the resulting score. Unlike the usual adaptive test where the computer only has access to difficulty and discrimination, if the item pool has been calibrated with a cognitively oriented model then the computer also has access to this information. To use word frequency again as an example, it is known that the vocabulary level of words that make up an analogy determines, in part, its difficulty. By focusing on difficulty alone as a means of selecting items we may end up choosing items that differ widely in their vocabulary load, in effect making the adaptive test for that individual a vocabulary test. Therefore, the ideal algorithm would insure that in addition to controlling difficulty it would insure that the blend of the different components of difficulty is maintained from one examinee to the next. This is possible if the item pool has information on the cognitive attributes of the item.

Scope of the SAT

It can be reasonably concluded from the foregoing that the SAT is "cognitively sound." That is, the SAT taps dimensions of human variability that are considered important by cognitive psychologists. Granted that we do not need to subtract anything from the SAT, does cognitive psychology have any suggestions for how to expand the SAT? At the recent Wakefield conference, two prominent cognitive theorists outlined theories which suggested ways in which the SAT could be expanded. Gardner, for example, argued that there are many more dimensions of human variability than those measured by the SAT. One could hardly disagree with that statement. For example, Gardner postulates spatial ability as one of the many "human intelligences." Of course, spatial ability has long been known by psychometricians to be an identifiable dimension of individual differences (e.g., Lohman, 1979). In the context of an admissions test however, the key question is whether the additional predictor is sufficiently informative to warrant the added expense and student time. With respect to spatial ability, it appears that the answer is no. Indeed, the SAT at one point did include a test of spatial ability (Gulliksen, personal communication), but it was dropped, apparently for lack of any predictive contribution. We do not mean to suggest that additional dimensions should not be considered, or in the case of spatial ability, reconsidered. Rather, so long as the purpose of the instrument is to help in the admissions process the incorporation of additional measures needs to be justified in terms of their contribution to admissions decision.

Research Implications

One more implication that may be drawn from this review is that while no changes to the content of the SAT appear to be necessary, additional research is needed to realize the programmatic implications just discussed. Except for the work reported on TSWE none of the existing research focuses on SAT items. This is especially true of quantitative items on which virtually no research, focused on processing models and determinants of difficulty, has been performed. In the verbal area the sentence completion item has also been totally ignored. Yet, ironically, the sentence completion item appears to be perhaps the most efficient of all verbal item types (Dorans, personal communication).

Cognitive Psychology and the SAT: A Review of Some Implications

Introduction

The content of the SAT has changed very little since its inception. Peter Loret's (1960) history of SAT item types provides a valuable chronological account of the introduction and removal of different item types. From the start there have been two major content areas, verbal and quantitative, and this is true today as well. The separation into verbal and quantitative parts was the result of Brigham's application of Spearman's factor analytic methods to early SAT forms (Donlon, 1984, p. 134). Although it may not appear so, even in retrospect, the division of the SAT into these two components is a profound statement about the organization of developed abilities, and as cognitive a statement as could be made today by any leading cognitive theorist. Nevertheless, the revival of cognitive psychology in the last two decades has radically changed the nature of psychological theorizing and concomitantly the view psychologists have of human thought (cf. Gardner, 1985). Perspicacious psychometricians, like John B. Carroll, appreciated early the impact that cognitive psychology would have on psychometrics (e.g., Carroll & Maxwell, 1979). The aim of this paper is to inquire whether the organization of the SAT needs revising, whether its scope needs to be enlarged, or the interpretations of scores modified, and whether the development of the test itself can be aided by results from cognitive psychology. The goal of this paper is to attempt answering those questions.

Background and Scope of the Paper

The current SAT verbal section consists of four item types: antonyms, analogies, sentence completion, and reading comprehension. The antonyms have been used since 1926, the only change occurring in 1953 from a six-option item format to a five-option one. Analogy items have been used in their present form since 1944; sentence completion has been used since 1946. Finally, reading comprehension has been used in its present form since 1946.

There has been a little bit of evolution in the quantitative section. Currently, there are two item types, the "regular math" item and quantitative comparison. Although the content of the item itself has evolved to reflect changes in the curriculum, the "regular math" item has been in its present form since 1942. The major change in the quantitative section has been the replacement of the data sufficiency item type with the quantitative comparison type. This change occurred in 1974.

The relative stability of the SAT has not adversely affected its usefulness as measured by its ability to predict first-year grade point average (Bejar and Blew, 1981; Donlon, T. E., 1984.) Nevertheless, there has been a substantial change in the mode and form of theorizing in psychology and education from a behavioristic to a cognitive perspective. Cognitive psychology, however, is not a "new" kind of psychology. It appears to be a novel approach only because it was preceded by a long period of behaviorism. Whereas behaviorism was concerned with modeling behavior as a function of the organism's experience, cognitive psychology is concerned with modeling behavior as a function of mental processes and structures.

To our knowledge no one wrote (and no committee requested) a paper on the implications of behaviorism for the SAT. Why then a paper on the implications of cognitive psychology for the SAT? Perhaps the best answer is that cognitive psychology, unlike behaviorism, is concerned with modeling performance on tasks not unlike those that appear in tests, including the SAT. Second, unlike behaviorism, cognitive psychology has shown an interest in individual differences. It is differences in performance on the SAT that form the basis for predicting success in college. Therefore, an understanding of why some students do better than others on the SAT is an important component of the validity of the SAT. Cognitive psychology, because of its affinity with individual differences research, may enhance our understanding of performance on the SAT as it currently stands and may suggest ways in which the SAT may be changed.

Despite the affinity between cognitive and individual differences research, it is valuable to mention at the outset a distinction between two types of tests in order to demarcate the scope of our paper. One type, of which the SAT is an example, may be called outcomes-oriented. The principal characteristic of this type of test is that, for scoring purposes. what matters is whether the the correct answer is chosen. Indeed, performance on the test is summarized as the sum of correct answers. By contrast, a process-oriented test is concerned equally with correct and incorrect answers. The purposes for which these two types of test are best suited are very different. A process-oriented test seems best suited for diagnostic assessment since as a result of the test it may be possible to prescribe a set of actions to improve performance. An outcomes-oriented test is best suited for ranking examinees as efficiently as possible. It is our belief that a test cannot easily have both functions. We equally believe that the SAT is not meant to be a diagnostic test. These assumptions have a direct impact on the scope of this paper. Specifically, even though cognitive psychology has much to offer to the design and construction of diagnostic instruments (Bejar, 1984) we will not discuss those implications in this paper. Instead, our focus is on the implications of cognitive psychology for the SAT as an outcomes-oriented test.

The outline for the rest of the paper is as follows. First, we will present a description of what cognitive psychology is. One of the methodologies that emerge from that discipline is "cognitive task analysis" and will be described next. We will then present applications of cognitive task analysis to the Quantitative, Verbal, and TSWE items to illustrate what we think is the most tangible implication of cognitive research for the SAT, namely a more detailed understanding of performance on the different item types that make up the SAT. Because there has been relatively little work on quantitative items we present a fairly elaborate description of cognitive task analysis but without any empirical findings. By contrast there has been a considerable amount of empirical work in the verbal domain, especially on verbal analogies, and thus we will be able to present some results. One impact that cognitive psychology is already having on psychometrics, and indirectly on the SAT, is a rethinking of the concept of construct validity. The next-to-last section of the paper discusses this rethinking of the concept and related SAT research. Finally, the last section discusses the programatic implications cognitive psychology has for the content, scope, and development of the SAT.

What is Cognitive Psychology?

Cognitive psychology is the scientific analysis of human mental processes and knowledge. Cognitive psychology views humans as "processors of information" and uses "analysis" as its major theoretical technique. Of particular relevance for the SAT are the analysis of processes used for successful learning and thinking in subject-matter domains and the analysis of knowledge acquired by learners in subject-matter domains.

The cognitive approach is based on a human/computer analogy. We can characterize computers in terms of their capacity characteristics (e.g., how many pieces of information can be held in a particular memory store?) and their processing characteristics (e.g., how many computations per second?). Analogously, we may be able to characterize the human/information processing system by using the same kinds of parameter.

The cognitive approach to the analysis of intellectual ability has the potential for extending and improving upon the traditional psychometric approach. While the psychometric approach tends to focus on "traits" or "factors," the cognitive approach seeks to describe theoretically important parameters of the human mind, some of which may be involved in students' intellectual performance on mathematical and verbal SAT items. While the psychometric approach tends to focus on the product of thinking, as indicated in final answers to questions the cognitive approach tends to focus on the process of the thinking; that is, the way that the student arrives at an answer. The psychometric approach is often concerned with the amount (or accuracy) of a student's knowledge, i.e., how much the student knows; the cognitive approach is concerned with the structure of a student's knowledge, i.e., what the student knows.

The cognitive approach is illustrated by Shepard's extremely influential research program on spatial cognition (e.g., Shepard & Metzler, 1971). This research program attempted to understand performance on a three-dimensional mental rotation task. An example of the kind of stimuli used in this research appears in Figure 1. While the figures are three-dimensional they have in common with spatial psychometric tests, such as those found in the Primary Mental Abilities (1938), their "rotational" nature.

- - - - - - - - - - - - - -

Insert Figure 1 About Here

- - - - - - - - - - - - - -

What was novel in Shepard's work was not the dimensionality of the stimulus but rather the attempt to describe in detail the mental <u>processes</u> and <u>representations</u> used by subjects to perform this sort of task. Shepard concluded that humans "mentally rotate" the figures to test congruence. He argued that the mental representation is an analogue one, rather than a propositional one, arguing on the basis of results that demonstrated the strikingly orderly relationship between the time to decide that the figures were, or were, not rotations of each other and angular disparity. That is, subjects would do the rotation in much the same way as one would rotate a picture of the figure. Although there is no unanimous support of this explanation (see Pylyshyn, 1973) what is more significant for our purposes is that this sort of explanation was quite different with respect to the explanation of psychological data offered other psychological models. In particular, this type of explanation addresses directly the thought processes needed to solve an item-like problem and is very different from psychometric accounts of performance on similar tasks, namely the postulation of spatial mental factors (e.g., Lohman, 1979).

Just as theorizing on spatial cognition relied on psychometric stimuli so has the theorizing on the verbal domain. Verbal analogies provided the prototypic task for the first theory to combine intelligence and cognition (Spearman, 1923). Spearman did not have available methods for studying empirically the information-processing components of verbal analogies; modern cognitive psychology, however, has devoted much effort to understanding verbal analogies.

Contemporary studies in cognitive psychology have consistently supported several general information processing components using diverse methodologies. These components include encoding (retrieving the word meaning from long-term memory), inference (educing relationships between word pairs), application of the relationship to a word, and response evaluation. Sternberg (1977a & b) applied mathematical modeling of response time to identify processing on verbal analogies and to study individual differences in aptitude. Other studies have combined protocol analysis with mathematical modeling (Pellegrino & Glaser, 1979; Whitely & Barnes, 1979).

Cognitive psychology has produced techniques for analyzing intellectual "hardware" (i.e., fundamental memory capacities and processing characteristics) and intellectual "software" (i.e., knowledge). The analysis of intellectual capacities and processing characteristics involves a description of the architecture of the human information-processing system. The information-processing system can be analyzed into a series of memory stores and processes for transferring information from one store to another. During the 1960s cognitive psychology emphasized a series of memory stores: sensory memory, a rapidly fading image of sensory experience; short-term memory, a limited-capacity store that can hold approximately five chunks of information but requires periodic rehearsal;

and long-term memory, an unlimited capacity store that holds information in an organized or meaningful mode but that loses information because of interference with retrieval routes. During the 1970s emphasis shifted to focusing on the processes for moving information through the system: attention moves information from SM to STM; rehearsal keeps information active in STM; encoding moves information from STM to LTM; and retrieval moves information frcm LTM to STM.

Analysis of knowledge, i.e., the contents of long-term memory, involves the major new focus of cognitive psychology during the 1980s. For example, Anderson (1985) has distinguished between declarative and procedural knowledge. Declarative Knowledge refers to knowledge about the world, such as the idea of what "the symbol + means to addition." Procedural Knowledge refers to knowledge about how to carry out some operation, such as the procedure for long division.

Another kind of knowledge that has received recent attention is Strategic Knowledge (Greeno, 1979) or what has been called "Meta-cognitive" Knowledge. Examples of Strategic Knowledge include planning, monitoring, adjusting one's processing for different goals, adjusting one's output for different audiences, and so on.

What is Cognitive Task Analysis?

One potential impact of cognitive psychology on the SAT is the application of cognitive task analysis to SAT items. Cognitive task analysis refers to specifying the cognitive capacities and knowledge that are required to successfully carry out a particular task, in this case, solving a test item. The two most common types of cognitive task analysis are the cognitive correlates approach and the cognitive components approach (Sternberg, 1979).

The cognitive correlates approach is based on the analysis of the "hardware" of the information-processing system. For any intellectual task, such as solving an SAT math or verbal item, the cognitive correlates approach asks, "How does the information-processing capacity of high-SAT students differ from that of low-SAT students?" The approach is to take a sample of students who score high on the SAT (or a subscale of the SAT) and a sample of students who score low, and compare these two samples on a series of tests of the information-processing capacities and characteristics. For example, Hunt (1985) has compared low- and high-verbal students on tests of the holding capacity of short-term memory, the speed of retrieving a verbal item from long-term memory, the speed of making a mental decision concerning verbal items, and so on. Rose (1980) has developed a battery of information-processing tests, and Carroll (1976) has shown how traditional psychometric factors may be used to tap characteristics of the information-processing system.

The cognitive-components approach is based on the analysis of the "software" of the information-processing system. For any intellectual task, such as solving an SAT verbal or math item, the cognitive-components approach asks, "What does someone have to possess in order to successfully perform on a particular SAT item?" The approach is to take an SAT item and logically determine the specific knowledge, processes, and strategies required to produce a successful answer. For example, Mayer (1985) has analyzed the solution of mathematical story problems into four main components: linguistic and factual (or declarative) knowledge is required for translation of each sentence of the problem; schematic knowledge (or knowledge of problem types) is needed for integration of the problem information into a coherent representation; strategic knowledge is required for developing and monitoring a solution plan; and algorithmic (or procedural) knowledge is needed for executing the plan. The product of this approach, with respect to the SAT, would be a list of components (i.e., specific knowledge) required for each SAT item.

The cognitive-components approach seems most useful for evaluating people's knowledge (i.e., achievement) in specific subject-matter domains, while the cognitive-correlates approach seems most useful for evaluating general intellectual ability. In addition, both approaches should be supplemented by analyses of the metacognitive strategies used by successful students.

Applying Cognitive Analysis to Verbal and Mathematical Tasks

Cognitive analysis techniques can be applied to verbal and mathematical tasks, including SAT items, with the goal of determining the underlying cognitive characteristics or components required to succeed on the SAT, that is, the identification of item attributes implicated in the difficulty of the items. A second goal is to modify items in such a way that performance on the test can be summarized in terms of the processes and structures postulated by the cognitive analysis of the item rather than as a global score. These two goals correspond to the distinction introduced earlier between outcomes-oriented and process-oriented tests. As we indicated, we do not feel that the SAT can reasonably be expected to serve both functions. Nevertheless, in this section we will present an analysis of how one quantitative item-type can be modified to better capture the process examinees are likely to follow to solve the item. We will then present analyses of verbal item-types which assume that the items can be left as they currently are, which we feel is a more realistic approach given the outcome orientation of the SAT.

Quantitative Items

As an example of how cognitive task analysis could be applied to a mathematics item from the SAT, consider the following sample problem from Taking the SAT:

> The members of a club decided to wash cars in order to earn money for the club. Each member of the club washed 3 cars and charged $2 per car. when they had finished, their receipts totalled $66, which included $6 in tips. How many members were in the club?
>
> (A) 9
>
> *(B) 10
>
> (C) 11
>
> (D) 20
>
> (E) 22

This item presents a word problem and asks the student to select the correct numerical answer.

A premise underlying the cognitive task analysis approach is that an item like the car wash problem requires several kinds of skills that can be distinguished and evaluated separately. A cognitive analysis of the skills required to solve this problem reveals that the successful student would need to know how to do the following: (1) translate the words of each sentence (or phrase) into another form of representation, (2) integrate the sentences (or phrases) into a coherent representation, (3) plan and monitor a solution strategy, and (4) carry out the operations in the plan. The first two skills involve techniques for representing a problem, while the last two skills involve techniques for solving a problem. The remainder of this section suggests how each of these skills could be evaluated using the standard SAT item format..

The first skill is translation of a problem sentence into another form of representation. Although this seems like an obvious skill, recent research on students' memory for word problems indicates that students make errors in comprehension of the problem sentences (Mayer, 1985). Even college students often fail to correctly generate an equation to represent a simple problem sentence. One way of testing this skill is simply to ask students to recognize paraphrases of the given information, such as the following question based on the car-wash problem.

Which of the following sentences is not true?

(A) Each member washed 3 cars.

(B) Each car was charged $2.

(C) The club took in a total of $66.

*(D) The club took in $6 in tips, in addition to $66 for washing cars.

(E) The total receipts of $66 included $6 in tips.

A related way of testing for translation skill is to ask the student to recognize paraphrases of the problem goals, such as the following question:

What are you being asked to find?

(A) how many cars were washed

(B) how much profit the club made

*(C) how many club members there are

(D) how much each member made

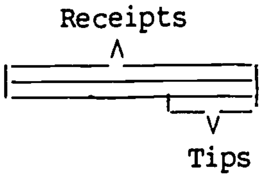
(E) who washed the most cars

Other techniques for evaluating translation skill include asking students to recognize equations or pictures that correspond to sentences in the problem. Two examples are given below:

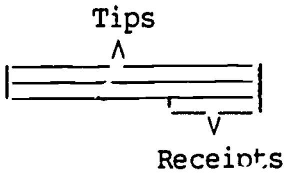
"The receipts totalled $66, which included $6 in tips."

Which of the following pictures corresponds to this statement?

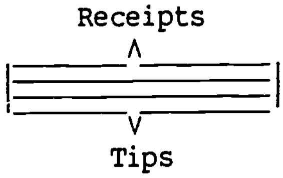
(A)



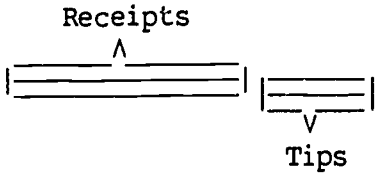
(B)



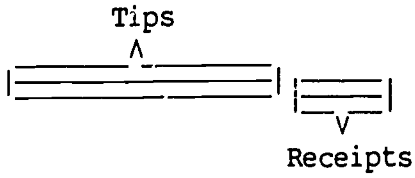
(C)



(D)



(E)



"Each member washed 3 cars and charged $2 per car."

Which of the following equations corresponds to this statement?

(A) each member's washing income = $2

(B) each member's washing income = $3

(C) each member's washing income = 2 * $3

*(D) each member's washing income = 3 * $2

(E) each member's washing income = $2/3

In addition to being able to represent each sentence in the problem, students must also be able to put the information together into a coherent structure. One possible evaluation technique for this problem-integration skill is to ask students to recognize relevant and irrelevant information in a problem. Since most mathematics items on the SAT contain only relevant information, it may be necessary to revise some items. For example, the car-wash problem could be rewritten to include some irrelevant information, such as the following:

> The members of a club needed to raise $100 to pay off a club debt. The members of the club decided to wash cars to earn money for the club. Each member worked four hours. Each member of the club washed 3 cars and charged $2 per car. When they had finished, their receipts totalled $66, which included $6 in tips. How many members were in the club?

Which numbers are needed to solve this problem?

(A) $100, 4 hours, 3 cars, $2, $66, $6

(B) $100, $2, $66, $6

(C) 3 cars, $2, $66

(D) 3 cars, 4 hours, $2, $66

*(E) 3 cars, $2, $66, $6

Other techniques for evaluating integration skill include asking the student to recognize an equation, number sentence, or picture corresponding, to the problem, such as the following (based on the original car-wash problem):

Which of the following expressions corresponds to this problem?

*(A) ($66 - $6)/(3 x $2) = ___

(B) ($66)/(3 x $2) = ___

(C) (3 x $2)/($66 - $6) = ___

(D) (3 x $2)/($66) = ___

(E) ($2/3) x $66 = ___

The foregoing items were example items for evaluating a student's skill at representing a problem, including translation skill and integration skill. In addition to these skills, a third skill is the ability to plan and monitor a solution strategy. One technique for evaluating planning skill is to ask a student to recognize subgoals in the problem, such as the following:

To answer this question you need to calculate:

(A) the age of each club member

(B) the tips received by each club

(C) the number of cars washed by each member

*(D) the total income from washing cars without tips

(E) the average time required to wash each car

In some cases, another evaluation technique for planning skill is to ask the student to recognize operations that must be carried out. For example, in the car-wash problem, a necessary operations question is:

To calculate an answer to this problem you must:

*(A) multiply, subtract, divide

(B) add, add, subtract

(C) divide, divide, divide

(D) multiply, multiply

(E) divide, multiply

Another technique for evaluating planning skill is to ask the student to draw conclusions concerning the results of a computation. For example, in the following problem the student is given all the numerical answers and asked only to make a conclusion.

You make the following computations:

3 x 2 = 6

66 - 6 = 60

60/6 = 10

10 x 3 = 30

6/10 = .60

Look back at the question. What is the answer?

(A) .6

*(B) 6

(C) 10

(D) 30

(E) 60

Finally, the fourth skill required to solve the car-wash problem is to carry out arithmetic and algebraic operations. For example, the following item is designed to evaluate arithmetic computation skill:

(66 - 6/(3 x 2) = ___

The correct answer is:

(A) 6

*(B) 10

(C) 20

(D) 30

(E) 60

This section has presented examples of how to evaluate four component skills for algebra word problems. Translation skill can be evaluated by asking students to recognize paraphrases of the problem given or goal, and to recognize pictures or equations corresponding to a sentence in the problem. Integration skill can be evaluated by asking students to distinguish relevant and irrelevant information, and to represent the problem as a number sentence, equation, or picture. Planning skill can be evaluated by asking students to identify problem subgoals, identify necessary operations, and draw a conclusion. Computational skill can be evaluated by asking students to identify the result of arithmetic problems.

Test items designed to evaluate component skills can provide more detailed information than items which focus only on the final answer. The inclusion of items that focus on problem representation and planning skills reflects a growing consensus among mathematics educators that instruction and evaluation should include emphasis on problem-solving process--such as how to represent a problem and how to plan a solution--as well as problem-solving product--such as getting the correct final answer (Mayer, 1985; National Council of Teachers of Mathematics, 1980). This trend towards evaluating students' skills in the process of problem solving is manifested in changes in standardized achievement testing programs carried out in school districts. For example, the California Assessment Program (1983) has been modified to include algebra word problems that focus on problem representation and planning rather than focus solely on getting the correct final answer.

Analysis of Verbal Items

Unlike mathematical items, verbal items have received much attention from cognitive psychologists interested in individual differences and from psychometricians interested in cognitive psychology. As a result of this interchange, the cognitive perspective has manifested itself in the form of psychometric response models that explicitly incorporate cognitive hypothesis into the parameters of the psychometric model.

This section presents the cognitive analysis for three item types, verbal analogies and paragraph comprehension and TSWE items. Because the word "model" will be used below in connection with "cognitive model" and "psychometric model" it is valuable to distinguish between these two types of model. By a cognitive model we will mean a psychological model of the item-solution process. By a psychometric model we mean a statistical model that, in addition to providing a description of the items, also allows us to make statistical inferences about items and examinees. A psychometric model in and of itself does not have psychological import. However, one of the implications of cognitive psychology is precisely to provide a psychological interpretation to generic psychometric models.

Verbal Analogies

Verbal analogies have a longstanding reputation as measuring intelligence. In fact, Spearman (1923) regarded analogies as the prototype of intelligent thought and as the best indicator of general intelligence, "g." Spearman's dual theory of cognition and intelligence was based on analogies.

Verbal analogies were the first item type to be studied intensively by cognitive component analysis (e.g., Sternberg, 1977a & b; Whitely & Barnes, 1979; Pellegrino & Glaser, 1979; Whitely, 1979). The various theories, although different in minor respects, generally agree on major processing components. Although most theories concern processing on analogies with three-term stems, rather than two-term stems as found on the SAT, they provide a good foundation for understanding SAT analogies.

Cognitive task analysis. The various theories postulate that verbal analogies are typically solved by a rule-oriented processing strategy in which each analogy term in the stem, in sequence, is processed as completely as possible. Consider the following analogy:

Cat : Tiger :: Dog : __________

1) Lion 2) Wolf 3) Bark 4) Puppy 5) Horse

Sternberg's theory of analogical reasoning postulated the following processing components:

1) Encoding; converting the word stimulus to a meaningful internal representation and subtasks that represent the various postulated processing components. For example, the word "Cat" would be represented as meaning a class of animals (i.e., felines) or a domestic animal.

2) Inference; postulating a relationship between the first two terms in the analogy. For example, the relationship for "Cat:Tiger" in the analogy above is "Tiger is a type of Cat".

3) Mapping; postulating a higher-order relationship to map the domain of the analogy (Cat:Tiger) to the range (Dog:?).

4) Application; applying the inference to the unmatched term to generate a solution to the problem. For example, the solution to the analogy must be such that "X" is a type of "Dog."

5) Justification; an optional component in which one of several alternatives, which is not the ideal response, is justified as the best answer.

Correct information must be obtained from each processing component for the analogy tc be solved. Importantly, specifying the components provides a means by which the test developer can manipulate item content to control processing. For example, the difficulty of the encoding process can be controlled by manipulating vocabulary level. Analogies with a high vocabulary level may measure primarily encoding, because inference and application cannot be validly attempted if the terms cannot be correctly encoded. Analogies with a lower vocabulary level can be constructed to measure primarily the inference process. For example, a relationship that is not salient between the stem word pair can be selected as the basis for the analogy. Relational saliency is defined as the accessibility of a given relationship for a word pair, apart from the analogy context. The saliency of a specified relationship for a word pair can be objectively measured prior to constructing analogies by indices such as mean response time elapsed before the relationship is inferred or by the mean number of relationships inferred prior to the target relationship.

A complexity that must be added to this theory presented above is how the components are processed. For exampie, are the various terms processed in an exhaustive or self-terminating fashion? This is an area of controversy between theories. Sternberg's (1977a & b) data indicated that inference is an exhaustive process, such that all possible word-pair relationships are inferred on the first exposure to the two related terms (e.g., Cat:Tiger). Thus, the first inference "X is a type of Dog" would not falsify either "Puppy" or "Wolf." Since only one can be selected, the examinee must then try the next inference on his or her list, such as "X is a type of non-domestic Dog."

For the difficult analogies that appear on psychological tests, research has indicated that inference is not exhaustive. The word-pair relationship that is inferred is influenced by the other terms in the analogy (Whitely & Curtright, 1979).

Processing strategies. The theory presented above is still not sufficiently complex to fully describe the different strategies examinees may use in solving the item. Embretson, Schneider, and Roth (in press) found that strategies other than the rule-oriented strategy presented above were necessary to account for performance on psychometric analogies.

An information-processing path diagram shows the postulated strategies lead task solution. Figure 2 presents Embretson et al (in press) information-processing model for verbal analogy items. Figure 2 shows several strategies that can lead to problem solution: (a) a rule-oriented strategy, as presented above, (b) an associational strategy, in which an examinee selects an alternative by its associative strength to the unmatched term, without respect to the inference relationship (e.g., in the example above "Puppy" probably has the highest associative relations. p with "Dog"), and (c) random guessing. Additionally, a response elimination strategy (not presented on Figure 2) was also hypothesized. In the response elimination strategy, a partially correct rule is used to eliminate some distractors, followed by guessing among the remaining distractors.

- - - - - - - - - - - - - -

Insert Figure 2 About Here

- - - - - - - - - - - - - -

Within each strategy are the processing components, as shown in Figure 2. Embretson's components are more global than those postulated by Sternberg. They include rule construction (which would include Sternberg's inference, mapping, and application), response evaluation (which would include justification), and an application probability for the strategy. These components are noted as events along the upper path in Figure 2.

These strategies and components can be better understood by considering the subtasks that assess them, presented on Table 1. The association strategy was assessed by presenting the unmatched term and the alternatives as an item. Note that the related pair is omitted from the "item" so that a rule-oriented strategy cannot be applied to solve the item. The rule-oriented strategy was assessed by presenting two subtasks: "Rule Construction" and "Response Evaluation". In the Rule Construction subtask, the stem is presented and the examinee states the relationship to the unmatched term required for a correct solution. In the Response Evaluation subtask, the correct rule is given, and the examinee chooses from the response alternatives. The Rule Construction subtask was scored for both a completely correct rule, which would eliminate all distractors, and a partially correct rule, which would eliminate at least one distractor.

- - - - - - - - - - - - -

Insert Table 1 About Here

- - - - - - - - - - - - -

Figure 2 also defines a mathematical model of item difficulty from the probabilities associated with the components within each strategy. The probability of solving an item by the rule-oriented strategy, for example, is given by multiplying the probabilities of the two associated components and the application probability. The probability of solving the analogy is given by the sum of all (three) paths that lead to the correct response. That is, the solution probability is given by the sum of the three strategy probabilities. It is beyond the scope of this paper to dwell on the mathematics of interfacing a model of processing seconds that are portrayed in Figure 2 to a psychometric model. The interested reader is referred to Embretson, 1985b.

Paragraph Comprehension Items

Paragraph comprehension items are an important item type because they are commonly used to measure both verbal aptitude and reading achievement. As yet no cognitive analysis of SAT reading comprehension item is available. However, Embretson and Wetzel (in press) studied paragraph comprehension items that had been developed to measure verbal aptitude on the Armed Services Vocational Aptitude Battery (ASVAB). The ASVAB paragraph comprehension items consist of a short paragraph, followed by a question about the text and four alternatives. Although the calibrations of cognitive characteristics are specific to the ASVAB, the model that they developed can be used to study paragraph comprehension items on other tests as well, including the SAT.

Cognitive task analysis. The theoretical model that was used to calibrate multiple-choice, paragraph-comprehension items is given in Figure 3 (Embretson & Wetzel, in press). It can be seen that two general components are postulated, (1) text representation and (2) decision. The text representation process consists of encoding the word meaning and coherence processing to link the words into meaningful propositions. The decision process also consists of encoding and coherence processes to understand the meaning of the alternatives. Additionally, however, the alternatives must be mapped to the text and evaluated for truth status.

Figure 4 presents a further elaboration of evaluating the truth status of the alternatives. Previous literature (Pellegrino & Glaser, 1979; Whitely & Barnes, 1979) had suggested that evaluating truth status consists of two distinct stages, falsification and confirmation, as shown in Figure 4. In falsification, an examinee attempts to find textual information that will falsify each alternative. If more than one alternative remains after falsification processing, then the examinee attempts to find information in the text that confirms the remaining alternative. If only one alternative remains after falsification, confirmation is not attempted.

- - - - - - - - - - - - - - - - -

Insert Figures 3 and 4 About Here

- - - - - - - - - - - - - - - - -

Previous research had indicated several stimulus features that influenced each of these processing components. Text representation is influenced by the density of various types of proposition and the arguments (Kintsch & van Dijk, 1978), word frequency (Kucera & Francis, 1967) and the percentage of content words (Drum, Calfee & Cook, 1981). The decision process is influenced by the falsifiability of the distractors (Sternberg, 1977a & b), the confirmability of the correct answer, the word frequencies of the alternatives (Drum, Calfee & Cook, 1981), percentage of relevant text, and the level of reasoning required to evaluate the alternatives (Anderson, 1972). To study the role of text attributes on reading comprehension, performance items were scored on the stimulus complexity factors for both the text representation and decision. Table 2 presents a list of the factors, along with their correlations with item difficulty and their weight, standard error, and significance.

- - - - - - - - - - - - - -

Insert Table 2 About Here

- - - - - - - - - - - - - -

It can be seen that text representation difficulty is significantly increased by a high density of modifier propositions (e.g., adjectives and adverbs in the text) and content words, but by a low density of connective propositions (i e., explicit links between phrases such as "due to," "in order to," etc.). Embretson and Wetzel (in press) interpret these findings as suggesting that modifier propositions burden the coherence processes, as more parsing is required, while connective propositions, in contrast, facilitate coherence processing since less inference is required to link propositions together.

Table 2 also presents the stimulus factors that influence the decision. It can be seen that the percentage of relevant text for confirming the key as correct, the falsifiability of the distractors, the confirmability of the correct alternative, word frequency for distractors, and level of reasoning to confirm the correct answer all significantly influence item difficulty. The results indicate that item difficulty is high when text mapping is hard (i.e., high percentage of relevant text), distractors are not easily falsified, the correct answer is not easily confirmed, low word frequencies exist for the distractors, and inductive or deductive reasoning is required in the decision.

Another way to see the cognitive components in the selection of reading comprehension items is to examine the relationship between components. Figure 5 plots the individual items on the difficulty of the text representation and the decision components. It can be seen that text representation and decision difficulty are relatively uncorrelated in the ASVAB item sets. Thus, they can be regarded as relatively independent sources of cognitive complexity in the items. This implies that the test developer can construct items that are difficult on text representation, decision, or both, depending on the goal of testing.

- - - - - - - - - - - - - -

Insert Figure 5 About Here

- - - - - - - - - - - - - -

An Application to the TSWE

Although the TSWE is not actually part of the SAT it is administered together with it. Relatively little work, from a cognitive perspective, has been done on these item types. The work to date has been oriented towards an accounting of item difficulty based on the syntactic characteristics of the items (Stabler, 1986; Bejar, Stabler & Camp, 1987). Stabler (1986) reviewed a variety of syntactic factors that have been shown to affect sentence comprehension. That review was used in lieu of the task decomposition step as a means of postulating item attributes that may account for an item's psychometric characteristics.

The idea that syntactic complexity influences sentence comprehension goes back at least to Miller's Derivational Theory of Complexity (Miller, 1962), which was based on Chomsky's Standard Theory (Chomsky, 1965). According to that theory a sentence must first be transformed into its "deep" form and then transferred to the semantic module where understanding actually takes place. Miller's theory predicted that the complexity of the transformations required to put a sentence into its deep form would be an important factor in comprehending the sentence.

The empirical tests of the theory were not totally successful. Moreover, the Standard Theory has essentially been discarded. According to perhaps the most definitive review of the Standard Theory (Fodor, Bever, & Garrett, 1974):

> "There seems no serious doubt that structural descriptions are, in this sense, psychologically real: they specify at least some of the descriptions under which linguistic messages are interpreted and integrated...the experimental evidence for the psychological reality of deep and surface trees is considerably stronger than the experimental evidence for the psychological reality of transformations" (pp. 273-4).

The Bejar-Stabler-Camp investigation (1987) focused on one structural consideration, namely the depth of the parse tree (cf. Yngue, 1960; Frazier, 1985). Specifically, the maximal depth of the sentence, the depth at which the error occurred, and the length of the sentence were investigated as possible determinants of difficulty of the items as estimated on a representative sample of the SAT test-taking population. The underlying psychological hypothesis is that depth is an indicator of the complexity of the sentence on which the item is based, and therefore should be related to item difficulty.

To illustrate, consider the two items that appear in Figure 6. The parse tree associated with these sentences appears in Figure 7. The length is computed simply as the number of words in the sentence. The depth is the number of nodes from the top down to the word level through the longest possible path. The depth of the error is the number of nodes from the top to the word where the error occurs. The length, depth, and depth of error for the two trees corresponding to the sample items are indicated in the figure, along with the difficulty value associated with each item.

- - - - - - - - - - - - - - - - - -

Insert Figures 6 and 7 About Here

- - - - - - - - - - - - - - - - - -

The results of the Bejar-Stabler-Camp investigation suggested that both depth and depth of error contributed to the prediction of difficulty, although in different ways. Depth, as one might expect, increases difficulty. Depth of error, on the other hand, appears to have a facilitating effect. At first sight one might think that the deeper the error the harder it would be to detect, and therefore in a prediction equation depth of error would have a positive weight. As it turned out, depth of error had a negative weight, which suggests that the deeper the error, the easier the item. Thus, rather than making the item harder, placing the error deeper in the sentence makes the item easier. Bejar and Stabler suggested that to the extent the parse tree is a valid representation of the mental processes necessary to comprehend the sentence, the deeper the error occurs, the more noticeable it is since it would occur at the point where a failure to find the anticipated word is, in fact, more noticeable.

Summary

Reviewing applications of cognitive psychology to the different item types we can see that one implication of a cognitive perspective is to modify the items so that the cognitive process is more evident. This was illustrated for one quantitative item type. The applications to the verbal items illustrated a rather different approach. Instead of modifying the items so that the cognitive processes are more explicitly measured, existing items are examined to see how much they require from the different postulated processes for a correct solution with the aim of accounting for the difficulty of the item. This, in a nutshell, is the most direct mechanism through which cognitive considerations can be incorporated in a psychometric model and therefore the mechanism through which cognitive psychology can directly affect the SAT. This has broad implications for what constitutes validity as well as for other programmatic concerns. In the following section we will revisit the concept of construct validity in light of cognitive psychology, and in the final section we will draw some programmatic implications of our scrutiny of the concept.

Test Validity

The validational evidence in favor of the SAT is primarily its ability to predict first-year college grades, but several empirical studies of the SAT and related tests have sought to clarify its construct validity, (Donlon, 1984, chapter 7). Studies on factor structure (Coffman, 1966), developmental trend consistency of ability (Hilton, 1979; Jones, Burton, & Davenport, 1982), susceptibility to coaching (Messick & Jungeblut, 1981), and relationship to more specific aptitudes (French, 1957).

Although the concept of construct validity (Cronbach & Meehl, 1955) has had major impact on research that supports the quality of aptitude tests. The nature of psychological theorizing has changed substantially since the original article on construct validity (Cronbach & Meehl, 1955). The current strength of the cognitive perspective has led psychology from functionalistic theories to structuralistic theories. More specifically, psychology now emphasizes explaining performance on the basis of the systems and subsystems of underlying processes rather than identifying antecedent/consequent relationships. Cronbach and Meehl's emphasis on building theory through the nomological network, which contained primarily antecedent (test score) to consequent (other measures) relationships, can be viewed as a functionalistic approach. The major problem faced by the construct validation researcher in accordance with Cronbach and Meehl's conceptualization was to determine the constructs that accounted for variance in test performance.

The application of methods from cognitive psychology to test items has made possible the decomposition of performance into underlying components, strategies, and knowledge structures. These methods often involve a mathematical model of task performance to estimate the impact of the various postulated components on item difficulty. As a result it is possible to postulate psychometric response models that specify the relationship of ability to various item characteristics. The estimation of ability in the context of these models replaces the test score and its variance and covariances as the major focus. Thus, the construct validation emphasis of accounting for variance in test scores seems irrelevant. Of course, it is not. The focus on covariation in construct validation research addresses the important issue of _response consistency_ (Bejar & Yocom, 1986; Messick, 1981). An alternative focus, and one that underlies recent psychometric thinking, [illegible] _response effort_. That is, just as we have insisted that test scores must correlate in certain ways with other measures in order to understand what the test measures, we must now, armed with knowledge about theories of cognition, insist on an understanding of what makes items easier or harder.

The revision of construct validity suggested by cognitive theorizing has led to a major reformulation of the validation process consisting of two stages: construct representation and nomothetic span (Embretson, 1983). This reformulation can be viewed as the culmination of debates on the role of structure and function in individual differences psychology (e.g., Messick, 1972; Carroll, 1972.) In Embretson's reformulation, a construct is a theoretical variable that is a source of individual differences (although it need not be.) Construct-representation research seeks to identify the theoretical mechanisms that underlie task performance by cognitive task analysis methods. That is, the component processes, strategies, and knowledge structures that underlie performance identify the construct(s) that is/are involved in the task.

Nomothetic-span research, in contrast, concerns the utility of the test for measuring individual differences. It refers to the span of relationships between the test score and other measures. Nomothetic span is supported by the frequency, magnitude, and pattern of relationships of the test score with other measures.

The bulk of the validity evidence for the SAT is of the nomothetic type. There is far less research on construct representation. Some examples have been presented in this paper for verbal analogies and reading comprehension and TSWE items. Further research is obviously needed, especially with respect to the mathematics section.

In Cronbach and Meehl's conceptualization of construct validity, construct representation is supported by the same data as nomothetic span. That is, the correlations of individual differences on the test with other measures both define the construct and determine the quality of the test as a measure of individual differences. Embretson's structuralistic conceptualization of construct validity has qualitatively different types of data to supp,rt construct representation and nomothetic span. The former is supported by data on how task variations influence performance, as in cognitive experiments, while the latter is supported by individual-differences correlations.

The relationship of construct representation to nomothetic span is elaborated by correlational research in which individuals are measured on the components, the test total score, and external measures. Figure 8 shows a conceptual model of the relationship of the underlying cognitive constructs to ability and the nomothetic span of the test.

Ability is postulated to be an intervening variable in Figure 8, meaning that it is an inductive summary that is completely explained by a weighted combination of the components. These components, in turn, account for the validity of the test. That is, the components explain the validity of the test in that a weighted combination of them can replace ability to describe the nomothetic span of the test. Thus, ability or developed ability as in the case of the SAT is merely a convenient referent for a particular combination of cognitive variables with a certain pattern of nomothetic span.

- - - - - - - - - - - - - -

Insert Figure 8 About Here

- - - - - - - - - - - - - -

Research on SAT relevant to current views on validity. A few studies on the SAT are relevant to information processing. The stimulus content of the SAT antonym items was examined by correlating standard word frequency with item difficulty (Carroll, 1980). The multiple correlation of stem and key-word frequencies with item difficulty was .80, thus showing good prediction by the stimulus properties of the individual words. Chall (1977) studied the reading-comprehension items over 28 years of the SAT, finding that readability and vocabulary level had decreased but that amount of text appropriate to answering the questions had increased. Inconsistent trends were found for the level of comprehension required to answer the questions, as measured by Bloom's (1956) taxonomy.

Processing on the SAT antonymns was studied by Connolly & Wantman (1964) using protocol analysis ("Think aloud"). They found support for an associative process, as the protocols were characterized by the examinee's "feeling" that an alternative was appropriate or not.

Although these studies represent a beginning for examining cognitive processing, they obviously fall short of a process model. For example, the initial support of Connolly and Wantman for associative processing has been succeeded by mathematical models of antonym items that predict item performance from the connotative and defining properties of the words (Sternberg, 1985). Subsequent research, including that presented in the previous section, provides stronger support for the nature of processing and the studies have the further advantage of providing a means to anticipate the difficulty of new items before empirical tryout.

Programmatic Implications

In the previous pages we have discussed what cognitive psychology is and have illustrated its applications to several item types that appear in the SAT and TSWE; we have also discussed the reconceptualization of construct validity, which has been catalyzed by developments in cognitive psychology. In this final section we would like to discuss at some length the programmatic implications for the SAT. Additional issues, such as the administration of the SAT by computer, will also be discussed under appropriate headings.

Test Content of the SAT

It would have been reasonable to anticipate at the beginning of this paper that cognitive psychology would have concrete suggestions for the content of the SAT. It seems that this is not the case, at least with respect to the verbal section. There has been far less research of the quantitative domain by both cognitive psychologists and cognitively oriented psychometricians. Thus this lack of suggestion applies primarily with respect to the verbal section.

If any changes were to be made to the SAT on account of cognitive psychology they are not likely to involve *removing* the item types currently used, because performance on those items is widely acknowledged to be a phenomenon well within the scope of interest of cognitive psychology. Moreover, there appears to be no obvious way in which cognitive psychology research can directly improve the measurement of developed ability in verbal skills. For example, Hunt & Pellegrino (1984) have concluded with respect to the assessment of intelligence that the joint application of cognitive psychology and computers was not likely to "extend the range of evaluation if only because the simple vocabulary test is such a good predictor." Such positive assessments of the pragmatic value of current psychometric instruments are always reassuring, especially coming from leading cognitive psychologists. It is even more reassuring when some leading psychologists argue that the differences between psychometric and cognitive approaches are of method rather than substance. Again, in the context of intelligence assessment, Sternberg (1984) has argued that:

> It is not at all clear that they [cognitive psychologists] would have much to contribute to psychometricians by way of useful feedback regarding test content, because when cognitive psychologists have used reference measures at all for external criteria for their tasks and theories, they have used intelligence tests and subtests rather than the behaviors these tests were intended to predict (p. 47).

This lack of suggestions from cognitive psychology on how to modify the content of tests is probably not a shortcoming of cognitive psychologists, as Sternberg seems to suggest, but rather a poignant indicator that psychometrics and cognitive psychology are concerned with the same phenomena. As noted by Sternberg himself (1984): ...."I would argue that cognitive psychology has provided a valuable complementary way of investigating pretty much the same construct psychometricians have been studying all along" (p. 49, italics added).

Granted that both cognitive and psychometric approaches to individual differences are concerned with the same phenomena, their characterization of the students' knowledge and skills is not as we have pointed out throughout the paper. Because psychometrics often is concerned with ordering students on a continuum, it stresses "how much the student knows." Cognitive psychology, however, does not have similar pragmatic objectives, and instead focuses on a more psychologically motivated description of "what the student knows." This description is governed by theoretical considerations rather than pragmatic ones and may not be directly usable for a psychometric characterization of individual difference, at least with respect to standard psychometric models. Psychometric models such as Item Response Theory (e.g., Lord, 1980), which is now used for equating the SAT, have put psychometric inference on a solid statistical foundation but require that certain constraints be met that, at least at first sight, do not have psychological import. The challenge then, as we see it, is to blend psychometric and cognitive concerns harmoniously. We should take advantage of the framework provided by psychometric models in order to insure valid conclusions and, by examining the cognitive literature, in order to focus on the processes and representations used by examinees to solve test items.

This is the spirit of reconceptualization of construct validity discussed earlier. The blending of psychometric and cognitive ideas suggested by that reconceptualization was demonstrated by the examples of research on verbal analogies, reading comprehension, and TSWE items. The common denominator in those studies was an accounting of item characteristics, such as difficulty, based on a cognitive analysis. An understanding of item difficulty is perhaps the most important implication of cognitive psychology for the SAT since it affects most aspects of the test including test development, test validity, score interpretation, and test administration.

Implications for Test Development and Test Administration

Just as attention to cognitive psychology has suggested a major reformulation of the concept of validity, it may also significantly affect the test development process of existing tests, and in turn lead to more valid and efficient measurement methods. The impact on test development construction is twofold. On the one hand, the psychometric response models that are used to explain the data may themselves be partly motivated by some psychological theory of the response process. On the other, the process of creating items may itself be enhanced by input from cognitive psychology. That is, instead of creating and discarding items through an exclusively empirical item it may be possible to d̶sign items with specific psychometric characteristics. In this section we address these two possibilities.

In order to put things in perspective it is valuable to review briefly the state of the art in test development. A most informative account has been provided by Wesman (1971). Wesman termed the test item writing process a creative one. He listed as qualifications of item writers a "well developed set of educational values" and an understanding of the individuals for whom the test is intended. In his description, the writer must produce items that satisfy the specifications of the test plan. This description suggests that though the process is not haphazard, it is certainly idiosyncratic, since it is largely the item writer's responsibility to interpret what is wanted and to generate items that meet the prescription.

Nothing in that description suggests how to ensure that a given item measures the intended trait in a particular domain. Presumably, a good item writer will seek to do this precisely because he or she is a good item writer. Similarly, a good item writer should be able to evaluate the work of other item writers. Unfortunately, we are left without any explicit guidelines for generating and evaluating pertinent items. It is this lack of explicit guidelines that makes item writing more an art than a science. With the help of cognitive psychology those guidelines can be made explicit and as a result the item-writing process can move closer towards a science.

It is important to mention that systematic approaches to item writing have been proposed before. For example, Guttman and his associates have suggested procedures to systematize the creation of items (Guttman & Schlesinger, 1967). The approach requires that the facets or content categories of the achievement domains and the facets of the possible responses be clearly spelled out. According to Guttman and Schlesinger (1967), the benefits of the facet approach include the possibility of writing distractors having varying degrees of attractiveness or distractors representing different types of error. We have illustrated this approach with quantitative items. Although Guttman himself shuns cognitive interpretations of test performance, the facet approach lends itself well to such interpretations. Feldman and Markwalder (1971), for example, constructed a map-reading test in which the distractors within each item would be attractive to children at different stages of cognitive development. They found that, by and large, children responded in accordance with Piaget's model of cognitive development.

A facet-like approach to item construction, even if not motivated by a cognitive model, could serve as the basis for changing the orientation of the SAT from a dichotomous scoring one to a "partial credit" one. That is, instead of scoring items as right or wrong, examinees could be given credit proportionally to the quality of their response. An important by-product of this approach is the possibility of improving precision of measurement. This possibility, as a matter of fact, has a long history within psychometrics and was even tried with the GRE (Reilly & Jackson, 1973).

The bulk of these efforts have a strongly empirical flair, that is, more precise scores are obtained by weighting the different alternatives in terms of the ability of those choosing the alternative (e.g., Echternacht, 1976; Raffeld, 1975; Reilly, 1975) or in terms of the biserial correlation of the alternative with the total test score (e.g., Davis & Fifer, 1959). Others, however, have emphasized a more rational approach. Coombs, Millholland, & Womer (1956) designed a testing procedure in which students were instructed to identify as many alternatives thought to be incorrect as possible; de Finetti (1965) and Shuford, Albert, & Massengill (1966) have discussed testing procedures in which the responder is encouraged, by how the test is scored, to reveal his or her confidence in the correctness of each response option offered; and Feldman and Markwalder (1971) have devised weights based on Piagetian theory by which to discriminate among wrong responses.

In theory, these approaches should not only increase precision but also yield additional valid psychometric information. However, empirical research (see Wang & Stanley, 1970; Weiss & Davison, 1981, for reviews) has not consistently demonstrated the value of the additional information so obtained. Although the reliability of scores does seem to increase when these scoring methods are used, their validity does not always increase. Nevertheless, in light of developments reviewed in this paper and developments in psychometric theory (e.g., Bock 1972; Samejima 1969, 1972) oriented to polychotomous scoring it may be valuable to revisit the possibility of allowing partial credit for responses other than the key.

While the original motivation for polychotomous models cannot always be traced to a specific psychological model of item performance, another family of psychometric models have turned out to be far more amenable to psychological interpretation even when they retain a dichotomous orientation. These psychometric models are known as linear logistic models (Fischer, 1973). The basic feature introduced by this type of psychometric model and derivation of it (e.g., Embretson 1985a & b) is the notion that the difficulty parameter is explainable in terms of other more basic parameters. When these more basic parameters correspond to item attributes that arise from a cognitive task analysis, we are in effect linking the psychological theory of item solution with a psychometric model of performance according to that psychological theory. As we stated earlier it is through this linkage that cognitive psychology can have an impact on the SAT. An explication of this linkage for all the quantitative and verbal item types should be undertaken as a means of "updating" the validational evidence in favor of the SAT. A possible outcome of such studies for verbal analogies, for example, is determining the relative contribution of word knowledge to the solution analogies. Although some effect should be expected, ideally the bulk of the processing effort should be due to the demands of the items or analogical thinking.

Equatability

An important objective of the test developer is to insure that forms are comparable from year to year. This is accomplished by following closely a set of guidelines that specify, among other things, the range and distribution of item difficulty and discrimination. Independent of these statistical requirements the test developer must also follow strict content guidelines. An outcome of linking cognitive task analysis and psychometric modeling is that the test developer has a richer set of item attributes to work with. To the extent that these attributes are based on a valid psychological model of item solution, there is a chance of improving comparability of forms from year to year. In the absence of psychological item attributes the test developer relies on categorizations of items whose psychological foundation is not well documented. For example, analogy items are classified along a number of dimensions, including one that sorts the items into a concrete-abstract-mixed trichotomy. Although it is a sensible categorization, its origin is not known.

Predicting Item Difficulty

In addition to improving comparabilit, a cognitively oriented approach has the potential to help test developers to anticipate the difficulty of items. This would be of significant practical importance. Bejar (1983), for example, has reported that at least with respect to TSWE items there is considerable room for improvement in the prediction of item difficulty. On the other hand, the combination of the test developer's expertise with information about the cognitive demands of the items is a much better predictor of item difficulty than either source of information alone (Bejar, Stabler, & Camp, 987). Similarly, there is evidence from an ongoing GRE analogy study (Bejar & Enright, in preparation) that cognitive information and test developer expertise are much better predictors of item difficulty than either source alone.

The practical applications of being able to anticipate difficulty include the possibility of reducing the need for pretesting of items. This could come about in at least two ways. First, if estimates of difficulty for unpretested items are available it may be possible to more effectively choose which one to actually pretest on the basis of the gaps that may exist in the item pool. For example, if there are plenty of easy items, then obviously there is no point in pretesting items that are estimated to be easy. Second, if difficulty can be estimated successfully then clearly we have a handle on what makes the items harder or easier to begin with. That knowledge could be put to use while the item is being developed rather than after the fact. This is becoming more feasible as the entire test development process is more and more assisted by computers. To give a simple example, while composing an item the test developer could instantaneously look up the vocabulary load of the item as currently drafted and revise if necessary.

Test Administration

These implications apply to the existing SAT as well as to an SAT that could perhaps be delivered by computer. In fact it is in the context of computer administration that the implications of cognitive approach would have the most far reaching implications.

Many of the criticisms of the SAT seem to stem from the fact that it is a difficult test. Because computerized test administration often relies on item response theory (Lord, 1980) it is feasible through computer administration to adjust the difficulty of the test for each examinee. That is, instead of measuring every examinee with equal precision, which is the usual objective of adaptive tests, a computerized test could be designed, not in effect to be the most psychometrically efficient test, but rather a test that would be easier than under normal adaptive procedures. To be sure this would entail loss of precision of measurement, but the loss could in turn be compensated by lengthening the test.

An application of adaptive testing technology to make the test easier could proceed on its own, independent of any cognitive considerations. Of course, by combining adaptive testing technology with cognitive item analysis we can not only control the real and perceived difficulty of the test but also help to maintain the nomothetic span of the test, that is, insure the validity of the resulting score. Unlike the usual adaptive test where the computer only has access to difficulty and discrimination, if the item pool has been calibrated with a cognitively oriented model then the computer also has access to this information. To use word frequency again as an example, it is known that the vocabulary level of words that make up an analogy determines in part its difficulty. By focusing on difficulty alone as a means of selecting items we may end up choosing items that differ widely in their vocabulary load, in effect making the adaptive test for that individual a vocabulary test. Therefore, the ideal algorithm would insure that in addition to controlling difficulty it would insure that the blend of the different components of difficulty is maintained from one examinee to the next. This is possible if the item pool has information on the cognitive attributes of the item.

Scope of the SAT

It can be reasonably concluded from the foregoing that the SAT is "cognitively sound." That is, the SAT taps dimensions of human variability that are considered important by cognitive psychologists. Granted that we do not need to subtract anything from the SAT, does cognitive psychology have any suggestions for how to expand the SAT? At the recent Wakefield conference two prominent cognitive theorists outlined theories which suggested ways in which the SAT could be expanded. Gardner, for example, argued that there are many more dimensions of human variability than those measured by the SAT. One could hardly disagree with that statement. For example, Gardner postulates spatial ability as one of the many "human intelligences." Of course, spatial ability has long been known by psychometricians to be an identifiable dimension of individual differences (e.g., Lohman, 1979). In the context of an admissions test however, the key question is whether the additional predictor is sufficiently informative to warrant the added expense and student time. With respect to spatial ability, it appears that the answer is no. Indeed, the SAT at one point did include a test of spatial ability (Gulliksen, personal communication), but it was dropped, apparently for lack of any predictive contribution. We do not mean to suggest that additional dimensions should not be considered, or in the case of spatial ability, reconsidered. Rather, so long as the purpose of the instrument is to help in the admissions process the incorporation of additional measures needs to be justified in terms of their contribution.

Research Implications

One more implication of what may be drawn from this review is that while no changes to the content of the SAT appear to be necessary, additional research is needed to realize the programmatic implications just discussed. Except for the work reported on TSWE none of the existing research focuses on SAT items. This is especially true of quantitative items on which practically no research focused on processing models and determinants of difficulty has been performed. In the verbal area the sentence-completion item has also been totally ignored. Ironically, the sentence-completion item appears to be perhaps the most efficient verbal item type (Dorans, personal communication).

Summary

That the SAT has not changed fundamentally since its inception without diminishing its pragmatic utility suggests that the SAT is appropriate for its intended purpose. Nevertheless, the social and scientific milieu have certainly changed since the 1920s. This paper reviewed the implications of scientific developments accompanying the revival of cognitive psychology. We distinguished early in the paper between a process-oriented or diagnostic test and an outcomes-oriented test such as the SAT. Since the SAT does not aim to be a diagnostic test the implications of cognitive theories for test construction that would otherwise be applicable were not emphasized. Nevertheless, an analysis was presented based on quantitative items that illustrates how tests with a diagnostic orientation could be based on cognitive principles. Most of the paper, however, was devoted to an explication of the linkage of cognitive principles with psychometric considerations for an outcomes-oriented test such as the SAT. The essence of that linkage is an accounting of item difficulty. That is, an understanding of the differences in difficulty among items. This understanding, can be viewed as an additional requirement for construct validity which until recently focused almost exclusively on an accounting of the covariation of test scores in terms of "abilities." In other words, explaining covariation and item difficulty is now an equally important aspect of validation. Although work from this enlarged validational perspective on SAT items is limited, there is work based on similar items, especially analogy and reading comprehension, which suggests that the SAT will fare well when relevant studies are conducted. Despite the positive outlook, additional studies need to be conducted on the basis of this enlarged validational framework. Even though the predictive power of the test is not likely to increase significantly because of this research it is nevertheless essential as a means of realizing the programmatic implications discussed in the paper.

References

Anderson, J. R. (1985). The architecture of cognition. Cambridge, MA: Harvard University Press.

Anderson, R. C. (1981). How to construct achievement tests to assess comprehension. Review of Educational Research, 42, 145-170.

Bejar, I. I. (1984). Educational diagnostic assessment. Journal of Educational Measurement, 21(2), pp 174-189.

Bejar, I. I. (1983). Introduction to item response models and their assumptions. In R. K. Hambleton (Ed.), ERIBC monograph on applications of item response theory. Vancouver: Educational Research Institute of British Columbia.

Bejar, I. I., & Blew, E. O. (1981). Grade inflation and the validity of the Scholastic Aptitude Test. American Educational Research Journal, 18(2), pp 143-156.

Bejar, I. I., & Enright, M. E. (in preparation). An analysis of item features contributing to the difficulty of GRE Analogies. Princeton, NJ: Educational Testing Service.

Bejar, I. I., Stabler, E., Jr., & Camp, R. (1987). Syntactic complexity and psychometric difficulty: A preliminary investigation. College Board Final Report. Princeton, NJ: Educational Testing Service.

Bejar, I. I., & Yocom, P. (1986). A generative approach to the development of hidden-figure items (RR-86-20-ONR). Princeton, NJ: Educational Testing Service.

Bloom, B. S., (Ed.). (1956). Taxonomy of Educational Objectives. Handbook I. The Cognitive Domain. New York: David McKay.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika 37, 29-52.

California Assessment Program. (1983). Student achievement in California schools: 1982-33 annual report. Sacramento: California State Department of Education.

Carroll, J. B. (1980). Measurement of abilities constructs. In Construct validity in psychological measurement: Proceedings of a colloquium on theory and application in education and employment. Princeton, NJ: Educational Testing Service.

Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new structure of intellect. In L. B. Resnick (Ed.), The nature of intelligence. Hillsdale, NJ: Erlbaum.

Carroll, J. B. (1972. Stalking the wayward factors. Contemporary Psychology, 17, 321-324.

Carroll, J. B., & Maxwell, S. (1979). Individual differences in ability. Annual Review of Psychology, 603-640.

Chall, J. S. (1977). An analysis of textbooks in relation to declining SAT scores. Appendix to On further examination: Report of the advisory panel on the Scholastic Aptitude Test score decline, Willard Wirtz, chairman. New York: College Entrance Examination Board.

Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge, MA: MIT Press.

Chronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Coffman, W. E. (1966). *A factor analysis of the verbal sections of the Scholastic Aptitude Test* (College Board Research and Development Report 65-6, No. 17). Princeton, NJ: Educational Testing Service. (Also ETS Research Bulletin 66-30).

Connolly, J. A., & Wantman, M. J. (1964). An exploration of oral reasoning processes in responding to objective test items. *Journal of Educational Measurement*, *1*(1), 59-64.

Coombs, C. H., Millholland, J. E., & Womer, F. B. (1956). The assessment of partial knowledge. *Educational and Psychological Measurement*, *16*, 13-37.

Davis, F. B., & Fifer, G. (1959). The effect of test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement*, *19*, 159-170.

de Finetti, B. (1965). Methods of discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical Statistical Psychology*, *13*, 87-123.

Donlon, T. E. (Ed.). (1984). *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests*. New York: College Entrance Examination Board.

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

Echternacht, G. (1976). Reliability and validity of item option weighting schemes. *Educational and Psychological Measurement*, *36*, 301-309.

Embretson, S. (1985a). The problem of test design. In S. Embretson, (Ed.), Test design: Developments in psychology and psychometrics. New York: Academic Press.

Embretson, S. (1985b). Multicomponent latent trait models for test design. In S. Embretson (Ed.), Test design: Developments in psychology and psychometrics. New York: Academic Press.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. Psychological Bulletin, 93, 175-197.

Embretson, S., Schneider, L. M., & Roth, D. L. (in press). Multiple processing strategies and the construct validity of verbal reasoning tests. Journal of Educational Measurement.

Embretson, S., & Wetzel, D. (in press). Component latent trait models for paragraph comprehension tests.

Feldman, D. H., & Markwalder, W. (1971). Systematic scoring of ranked distractors for the assessment of Piagetian reasoning levels. Educational and Psychological Measurement, 31, 347-362.

Fischer, G. H. (1973). The linear logistic model as an instrument of educational research. Acta Psychologica, 37, 359-374.

Fodor, J. F., Bever, T. G., & Garrett, M. F. (1974). The psychology of language. New York: McGraw-Hill.

Frazier, L. (1985). Syntactic complexity. In D. R. Dowty, L. Karttunen, & A. M. Zwicky (Eds.), Natural language parsing (pp. 129-189). New York: Cambridge Press.

French, J. W. (1957). Validation of the SAT and new item types against four-year academic criteria. ETS Research Bulletin 57-4. Princeton, NJ: Educational Testing Service.

Gardner, H. (1985). *The mind's new science: A history of the cognitive revaluation*. New York: Basic Books.

Greeno, J. G. (1979). Cognitive objectives of instruction: Theory of knowledge for solving problems and answering questions. In D. Klahr (Ed.), *Cognition and instruction*. Hillsdale, NJ: Erlbaum.

Guttman, L., & Schlesinger, I. M. (1967). Systematic construction of distractors for ability and achievement test items. *Educational and Psychological Measurement*, *27*, 569-580.

Hilton, T. L. (1979). "ETS study of academic prediction and growth." *New Directions for Testing and Measurement, No. 2*, pp 27-44. San Francisco, CA: Jossey-Bass.

Hunt, E. (1985). Verbal ability. In R. J. Sternberg (Ed.), *Human abilities: An information processing approach*. New York: Freeman.

Hunt, E., & Pellegrino, J. (1984). *Using interactive computing to expand intelligence testing: A critique and prospectus* (Technical Report 84-2). Seattle: University of Washington.

Jones, L. V., Burton, N. W., & Davenport, E. C. (1982). *Mathematics achievement levels of black and white youth* (L. L. Thurstone Psychometric Laboratory Report #165). Chapel Hill, NC: University of North Carolina.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of test 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.

Lohman, D. F. (1979). *Spatial ability: A review and reanalysis of correlational literature* (Technical Report No. 8, Aptitude Research Project). Stanford, CA: Stanford University.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Loret, P. G. (1960). *A history of the content of the Scholastic Aptitude Test* (TDM-60-1). Princeton, NJ: Educational Testing Service.

Mayer, R. E. (1985). Implications of cognitive psychology for instruction in mathematical problem solving. In E. A. Silver (Ed.), *Teaching and learning mathematical problem solving: Multiple research perspectives*. Hillsdale, NJ: Erlbaum.

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin*, *89*, 575-588.

Messick, S. (1972). Beyond structure: In search of functional models of psychological process. *Psychometrika*, *37*(4), 357-75.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin*, *89*(2), 191-216.

Miller, G. A. (1962). Some psychological studies of grammars. *American Psychologist*, *17*, 748-762.

National Council of Teachers of Mathematics. (1980). *An agenda for action: Recommendations for school mathematics of the 1980's*. Reston, VA: NCTM.

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

Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. Psychological Bulletin, 80, 1-24.

Raffeld, P. (1975). The effects of Guttman weights on the reliability and predictive validity of objective tests when omissions are not differentially weighted. Journal of Educational Measurement, 12, 179-185.

Reilly, R. R. (1975). Empirical option weighting with a correction for guessing. Educational and Psychological Measurement, 35, 613-619.

Reilly, R. R., & Jackson, R. (1973). Effect of empirical option weighting in reliability and validity of an academic aptitude test. Journal of ducational Measurement, 10, 194-195.

Rose, A. (1980). Information-processing abilities. In R. E. Snow, P. Federico, & W. E. Montague (Eds.), Aptitude, learning and instruction. Hillsdale, NJ: Erlbaum.

Samejima, F. (1972). A general model for free-response data. Psychometrika, Monograph Supplement No. 18.

Samejima, F. (1969). Estimating latent ability using a response pattern of graded scores. Psychometrika, Monograph Supplement No. 17.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. Science, 171, 701-703.

Shuford, E. H., Albert, A., & Massengill, H. F. (1966). Admissible probability measurement procedures. Psychometrika, 31, 125-145.

Spearman, C. (1923). The nature of 'intelligence' and the principles of cognition. New York: Macmillan.

Stabler, E., Jr. (1986). Possible contributing factors in test item difficulty (RM-86-7). Princeton, NJ: Educational Testing Service.

Sternberg, R. J. (1985). *Beyond IQ: A triarchy theory of human intelligence*. Cambridge: Cambridge University Press.

Sternberg, R. J. (1984). What cognitive psychology can (and cannot do) for test development. In B. Hake (Ed.), *Social and technical issues in testing: Vol. 1*. Hillsdale, NJ: Erlbaum.

Sternberg, R. J. (1979). The nature of mental abilities. *American Psychologist*, *24*. 214-230.

Sternberg, R. J. (1977a). Component processes in analogical reasoning. *Psychological Review*, *31*, 356-378.

Sternberg, R. J. (1977b). *Intelligence, information-processing and analogical reasoning: The componential analysis of human abilities*. Hillsdale, NJ: Erlbaum.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.

Wang, M. W., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, *40*, 663-705.

Weiss, D. J., & Davison, M. L. (1981). *Review of test theory and methods* (Research Report 81-1). Minneapolis: University of Minnesota, Department of Psychology.

Wesman, A. G. (1971). Writing the test item. In R. L. Thorndike (Ed.), *Educational measurement*. Washington, DC: American Council on Education.

Whitely, S. E. (1979). Estimating measurement error on highly speeded tests. *Applied Psychological Measurement*, *3*, 141-159.

Whitely, S. E., & Barnes, G. M. (1979). The implications of processing event sequences for theories of analogical reasoning. *Memory and Cognition*, *1*, 323-331.

Whitely, S. E., & Curtright, C. A. (1979). *The role of problem representation in solving verbal analogies*. Presented at the annual meeting of the American Psychological Association. New York.

Yngue, V. H. A. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, *104*, 444-661.

Table 1

Subtask Set

Total Item	Cat : Tiger : : Dog: ______________
	(a) Lion (b) Wolf (c) Bark (d) Puppy (e) Horse
Association subtask	Dog
	(a) Lion (b) Wolf (c) Bark (d) Puppy (e) Horse
Rule Construction	Cat : Tiger : : Dog: ______________
	Rule ________?________
Response Evaluation	Cat : Tiger : : Dog: ______________
	(a) Lion (b) Wolf (c) Bark (d) Puppy (e) Horse
	Rule: A large or wild canine

Table 2

Complexity Factor Weights for Proposed Model

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	.22	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 **

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

Figure 1

Sample True and False Three-Dimensional Rotation Items

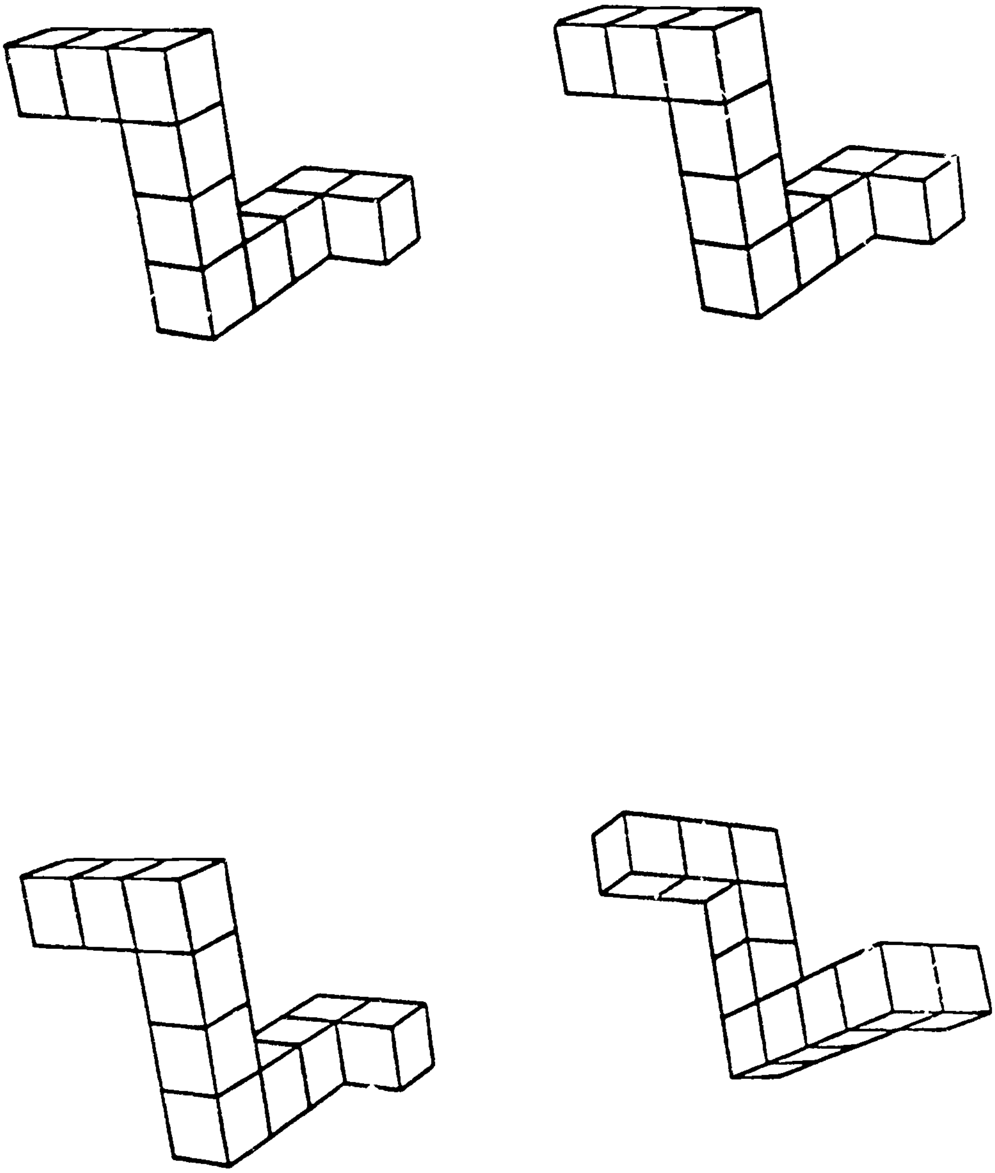


Figure 2

An Information-processing Path Diagram for Verbal Analogies

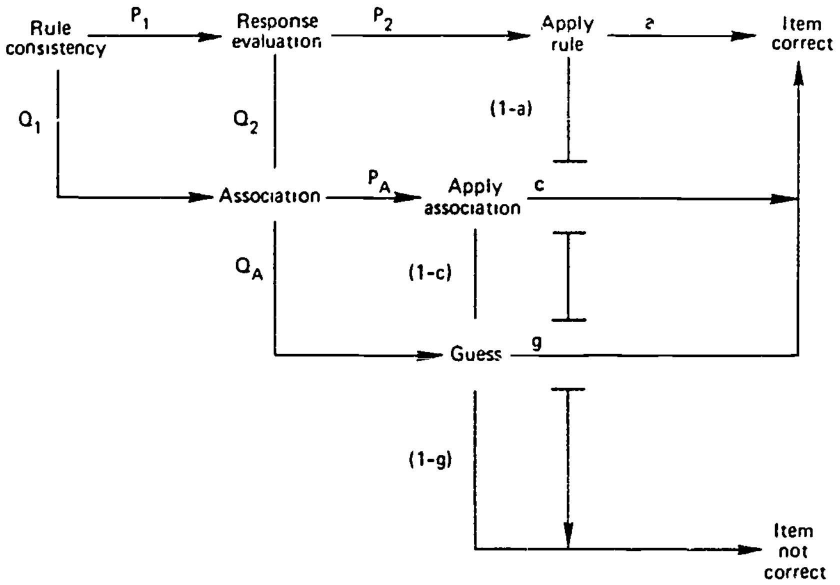


Figure 3

Information Processing Model of Paragraph Comprehension and Submodel for Evaluating the Truth Status of Response Alternatives

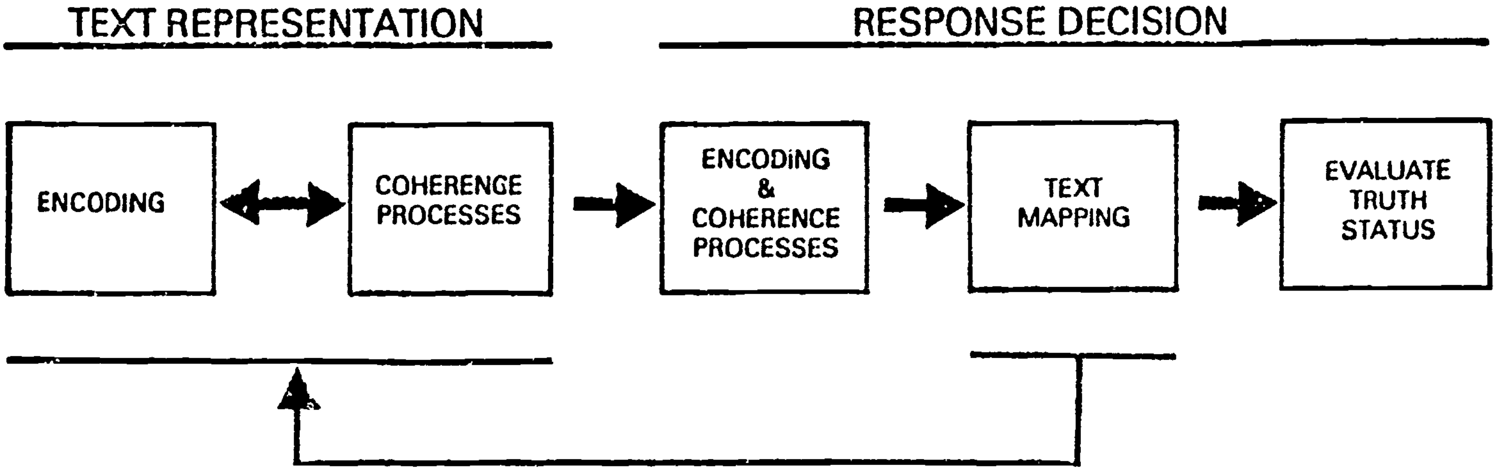


Figure 4

MODEL FOR EVALUATING ALTERNATIVES

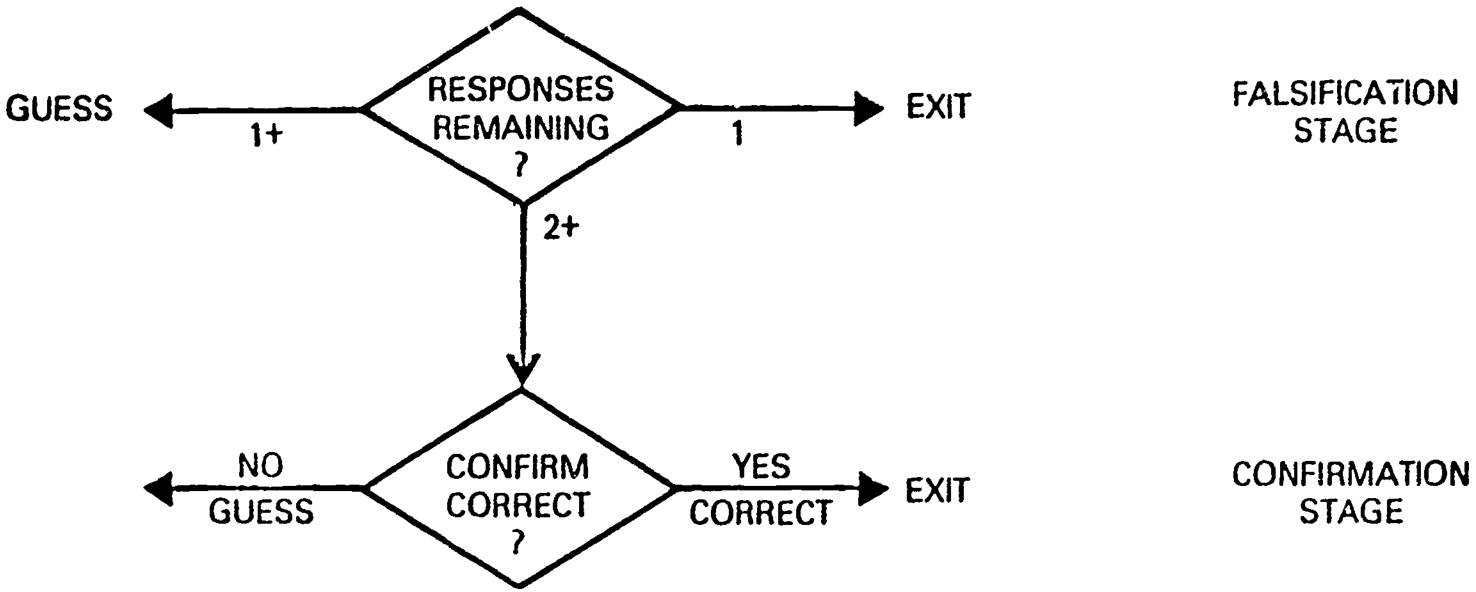


Figure 5

Relationship of Text Representation and

Decision Difficulty for ASVAB Item

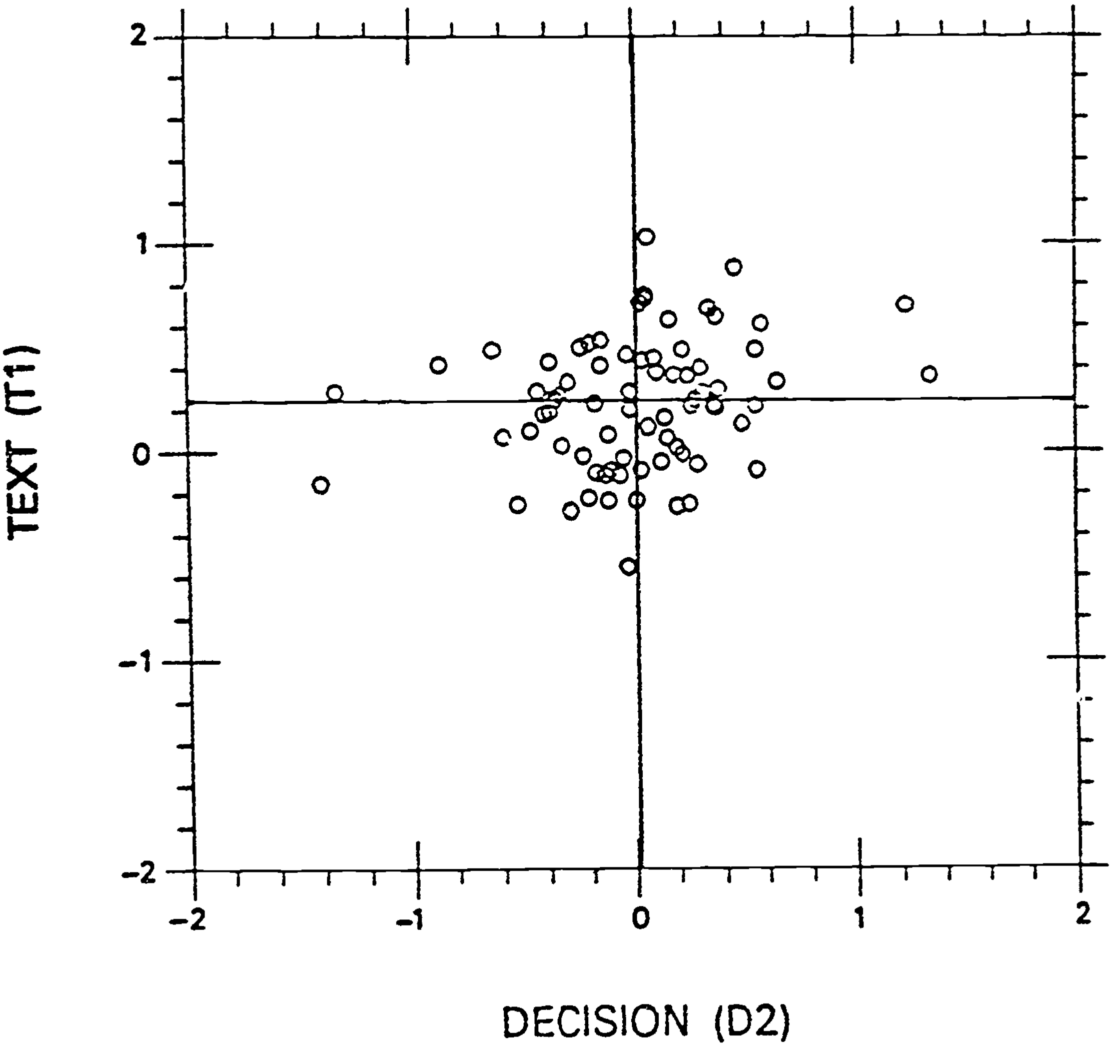


Figure 6

Sentences from Two TSWE Items

6-6. MARK SUGGESTED THAT THE MAJOR DIFFERENCE BETWEEN
A

HOGARTH'S ETCHINGS AND REMBRANDT IS IN THE VIEW
B C

EACH ARTIST HAD OF HUMANITY. NO ERROR
D E

6-20. EVERY GREAT INTELLECTUAL, RELIGIOUS, ECONOMIC,

OR SOCIAL DEVELOPMENT THAT HAS TAKEN PLACE IN
A B

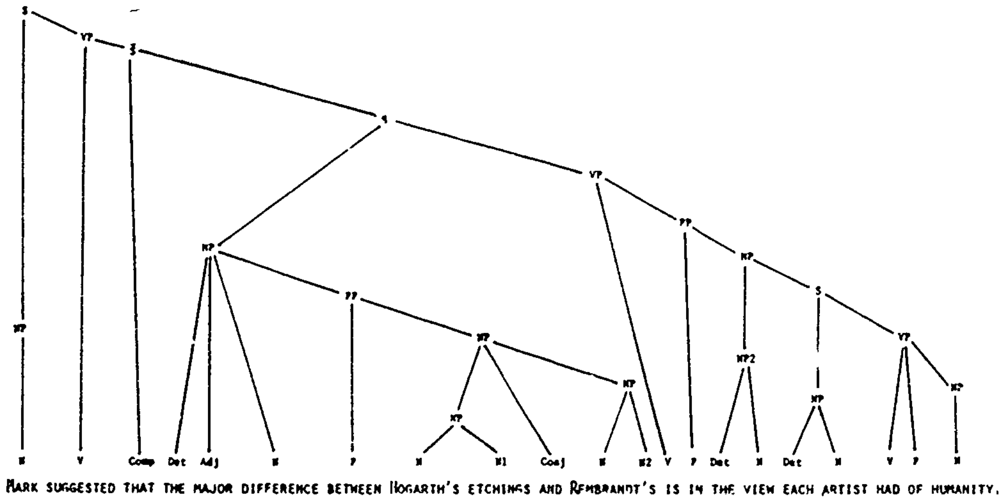
WESTERN EUROPE HAVE AFFECTED ENGLAND
C

SOONER OR LATER. NO ERROR
D E

Figure 7

Parse Trees for TSWE Items 6-6 and 6-20

Parse Tree for Sentence from Item 6-6 with Difficulty 11.9


S
VP
S̄
S
VP
NP
PP
PP
NP
S
NP
NP2
VP
NP
NP
NP
NP
N
V
Comp
Det
Adj
N
P
N
N1
Conj
N
N2
V
P
Det
N
Det
N
V
P
N
MARK SUGGESTED THAT THE MAJOR DIFFERENCE BETWEEN HOGARTH'S ETCHINGS AND REMBRANDT'S IS IN THE VIEW EACH ARTIST HAD OF HUMANITY.

Parse Tree for Sentence from Item 6-20 with Difficulty 6.5

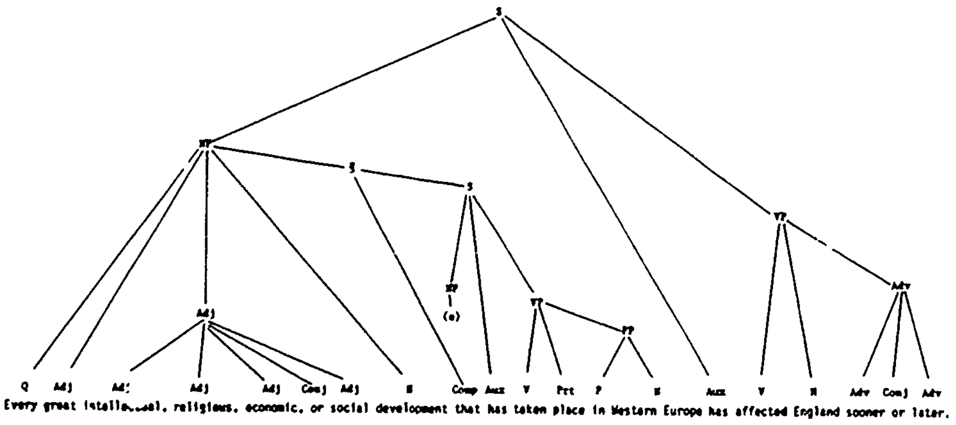
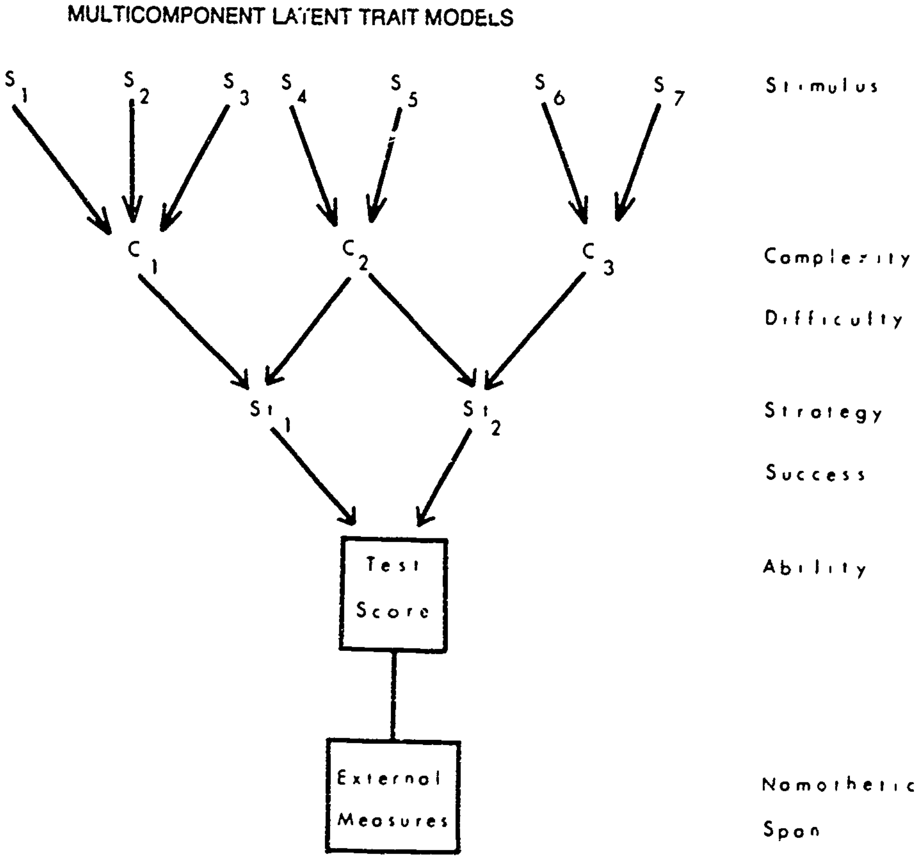

S
NP
S̄
S
VP
NP
(e)
VP
PP
Adv
Adj
Q
Adj
Adj
Adj
Adj
Conj
Adj
N
Comp
Aux
V
Prt
P
N
Aux
V
N
Adv
Conj
Adv
Every great intelle...al, religious, economic, or social development that has taken place in Western Europe has affected England sooner or later.

Figure 8

Relationship of Cognitive Variables to Test Score


MULTICOMPONENT LATENT TRAIT MODELS
S1
S2
S3
S4
S5
S6
S7
Stimulus
C1
C2
C3
Complexity
Difficulty
St1
St2
Strategy
Success
Test Score
Ability
External Measures
Nomothetic
Span