DOCUMENT RESUME

ED 224 840 TM 830 039

AUTHOR Quellmalz, Edys S.; Shaha, Steven
TITLE Cognitive Models for Integrating Testing and Instruction, Phase II. Methodology Program.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE 30 Nov 82
GRANT NIE-G-80-0112
NOTE 49p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; Cognitive Measurement; Criterion Referenced Tests; Elementary Secondary Education; *Instruction; *Mathematical Models; Norm Referenced Tests; *Task Analysis; Testing Problems; *Test Interpretation; Test Items; Test Theory
IDENTIFIERS Stanford Achievement Tests

ABSTRACT

The potential of a cognitive model task analysis scheme (CMS) that specifies features of test problems shown by research to affect performance is explored. CMS describes the general skill area and the generic task or problem type. It elaborates features of the problem situation and required responses found by research to influence performance. Stimulus features specify particular concepts and procedures necessary to answer a question and place limits on the structure and range of content or examples that can illustrate a problem. Response requirements indicate the mode as well as solution procedures and operations. The task content describes time constraints and the purpose, function, or audience of the task. CMS was used to classify and compare the contents of a Stanford Achievement Test and a criterion referenced test, both administered in 1982. Content analyses indicated the tests measure different aspects of reading and with different degrees of emphasis. Response patterns implied that CMS may be a promising tool for describing, analyzing, and interpreting test performance. The detailed partitioning of skill requirements used in the CMS appears to present a clear and interpretable test analysis tool. (Author/PN)

DELIVERABLE

November 30, 1982

METHODOLOGY PROGRAM: Cognitive Models for Integrating Testing and Instruction, Phase II

Edys S. Quellmalz
Study Director

Steven Shaha

Bruce Choppin
Project Director

Grant No.: NIE-G-80-0112
P3

Center for the Study of Evaluation
Graduate School of Education
University of California - Los Angeles

COGNITIVE MODELS FOR INTEGRATING TESTING AND INSTRUCTION

Edys S. Quellmalz
and
Steven Shaha

As studies of the development of subject-matter expertise identify features of cognitive tasks that distinguish among levels of performance, psychologists add their voices to the growing criticism of the kinds of problems and questions presented on achievement tests. Cognitive studies have been refining a research paradigm in which the performances of masters and novices are contrasted on problems differing in the complexity of information and procedures required to solve the task. When psychologists compare the kinds of tasks they develop to those presented in published tests and curricula, they find that many of these materials do not have sufficiently precise ways to describe and design test tasks and instructional activities or to analyze and interpret student test performance.

The issue underlying the rising tide of criticism of published achievement tests is their construct validity--whether the tests adequately define and sample the skill domain they purport to measure. Researchers maintain that the correlational techniques used by psychometricians to establish concurrent and predictive validity emphasize the metric and ignore the psychological bases of observed correlations; these researchers also find descriptive procedures for establishing content validity too superficial to distinguish among the requirements of tasks. Consequently, educators and psychologists from diverse disciplines are calling for a new form of test which designs

tasks and reports performance that is sensitive to the relationship between the features of a test problem and its required cognitive components --the stores of information and procedures the student must use to solve a task. (Glaser, 1981; Quellmalz, 1981; Sternberg, 1981.)

An example of this approach to analyzing the test task structures is a study by Bauman (1982) which challenges the construct validity of reading comprehension tests. He examined the linguistic structures of three standardized reading tests and found that the loose structure of reading passages allowed students to arrive at more than one legitimate answer. He also found that revisions of the text to clarify connections among ideas improved performance on some questions. In a similar vein, Langer (1981) reports on an extensive series of studies describing problems readers encounter in answering standardized reading test questions. The project drew upon linguistic and schema theory to develop a profile of "inconsiderate" test items that could mislead students' interpretations of a passage's genre, content, or semantic/syntactic structures.

To the extent that a task structure scheme could specify the stimulus features and response requirements of problems or tasks that affect skilled performance, such a scheme would be invaluable for designing test and instructional tasks and for interpreting performance on them. A detailed task structure scheme could guide the development of homogeneous pools of instructional or test problems and provide a basis for integrating testing with instruction.

The purpose of this study was 1) to cull from cognitive research the

task features that seem most important in determining performance, 2) to see how the items on existing tests represent these task features, and 3) to see if student performance differs on items with varying task features.

The study proceeded in two phases. In the first phase, we reviewed the growing body of learning research that points to particular features of problems or tasks that significantly affect performance. Based on this research we developed a cognitive model task structure scheme. We also specified two other task analysis schemes characteristic of those used to design and report objectives-based and standardized, general curricula tests. We then used these three task structure schemes to analyze a standardized and a criterion-referenced reading test.

The intent of the first phase of the study was to develop and refine the cognitive task structure scheme and to examine whether it provided distinct descriptions of tests' content. As we expected, the application of the cognitive model scheme yielded a fine-grained picture of the differences in the distribution of task features on the two tests. We also found that the detail captured by the cognitive model was glossed over in the more global objectives-based and general curricula schemes. In Phase I we established that a cognitive model task structure scheme identified differences among the kinds of tasks presented on tests -- task differences that cognitive research suggests would result in differences in student performance. The purpose of Phase II, therefore, was to move from descriptive to empirical comparisons. In Phase II we examined whether the clusters of items classified by the task structure scheme predicted patterns of performance.

The task structure schemes

In Phase I of the study, three task structure schemes were derived from an analysis of three approaches to designing tests and interpreting performance. The general curricula scheme presents the most global dimensions and is characteristic of many standardized tests. These dimensions include general descriptions of the skill (e.g., computation) and the content (e.g., fractions). The second scheme, the objectives-based task structure scheme, includes dimensions frequently used to develop and/or interpret scores on competency tests (Baker, 1974; Hambleton & Simon, 1980; Hively, 1974; Popham, 1978). In addition to a general skill description, the objectives-based scheme specifies stimulus attributes such as the range of concepts and content presented in test problems and response attributes including response mode (recognition and production) and rules for correct and incorrect responses.

The most detailed scheme is the cognitive model scheme which also describes the general skill area and the generic task or problem type. In addition, the cognitive model elaborates features of the problem situation and required responses found by research to influence performance. Stimulus features specify particular concepts and procedures necessary to answer a question and place limits on the structure and range of content or examples that can illustrate a problem. Response requirements indicate the mode as well as solution procedures and operations. The task context describes time constraints and the purpose, function, or audience of the task. Table 1 presents the components of a cognitive model task structure scheme.

Insert Table 1 here

METHOD

The second phase of the project investigated whether the three task structure schemes classified items into clusters that display distinct response patterns. The study addressed two questions. The first was whether any of the three levels of detail specified in the three task structure schemes yields more homogeneous patterns of performance and, therefore, provides a more stable, potentially useful picture of reading strengths and weaknesses. For example, is performance more consistent within a broad skill such as "inferential" questions, within subsets of inference questions differentiated by the locus of required information (within or beyond the test), or by the particular concepts such as mood, author's purpose or main idea?

The second question addressed the stability of performance across tests developed according to the different approaches: i.e., is performance consistent on sets of items testing similarly labeled skills such as inferential comprehension or opinion?

To address these questions, the study required test data at the individual and item levels. Interestingly, few school districts could retrieve test scores at this level because the test scoring and reporting procedures of test publishers and most districts produce and store only aggregate scores for subscales and test totals--a practice that does not permit districts to monitor test and item sensitivity to skill growth.

Table 1

Cognitive Model for a Task Structure Scheme

I. General Skill Area

A description of the task objective

II. Task Genre

A description of the problem type

III. Stimulus Features

Required stimulus features

Identification of the facts, concepts or principles targetted by the question or problem

Problem characteristics

Description of the problem form and substance, including:

Form - prose or symbolic

prose - topic familiarity, knowledge source, concreteness

structure - coordinate, subordinate, mixed

symbolic

elements - their type and number

Structure - explication of relations between elements

IV. Response Complexity

Description of the response mode

Description of the required operations

V. Context

Description of time limits,
purpose/function
audience

We did, however, find a district with individual and item level test data on sixth grade standardized and criterion-referenced reading tests. The cognitive task structure scheme was used to classify the passages and questions and student performance on the items in each cluster was then described and analyzed.

The tests

One of the tests was a standardized test, the Stanford Achievement Test, Intermediate Level II, Form A (SAT). The second test was a criterion-referenced reading test developed by the school district to assess its objectives. Approximately 640 students in the district had taken both of the tests in the Spring of 1982.

The SAT is a multipurpose test assessing student achievement in eleven areas including reading, math, social science, science, and listening. The reading comprehension test presents seventy-one questions about twelve passages that represent a range of discourse types. Reports of pupil performance to teachers present the total number correct, a scaled score, a grade equivalent score, a percentile rank and a stanine score.

The school district-developed criterion-referenced reading test presents seventy-six items for six reading passages. Of the seventy-six questions, twenty-six query students' comprehension of discourse, the rest relate to vocabulary, structural analyses, and reference skills. Scores are reported by objective.

The task structure scheme for reading

Terms drawn from tests and research on reading were used to apply the

generic cognitive model of task structure to the skill domain of reading comprehension. The general skill areas became literal and inferential comprehension and the general "content" areas became commonly tested features such as main idea, organization, detail, and figurative language. For this analysis, the particular feature comprising the "content" dimension were those specified in each test.

Figure 1 presents a "test tree" to illustrate the test item features referenced in a cognitive model, objectives-based or general curricula task structure scheme.

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

Insert Figure 1 here

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

The general curricula scheme references two "branches": the response requirements of the question and the reading level of the passage. For reading comprehension, standardized reading tests often report separate scores for literal and inferential questions; they do not report performance on various types of reading passages. Objectives-based reading comprehension tests report performance on more branches of the tree. They may report performance on literal and inferential questions and also performance on particular passage features such as mood or author's purpose. Some objectives-based tests specify the range of topics and length of passages, although score reports do not ordinarily reference these features.

The cognitive model task structure scheme specifies the most branches of the tree. A cognitive model specifies processes required by the question and additional features of the passage. Literal questions are divided into those requiring verification of information given verbatim in the

FIGURE 1

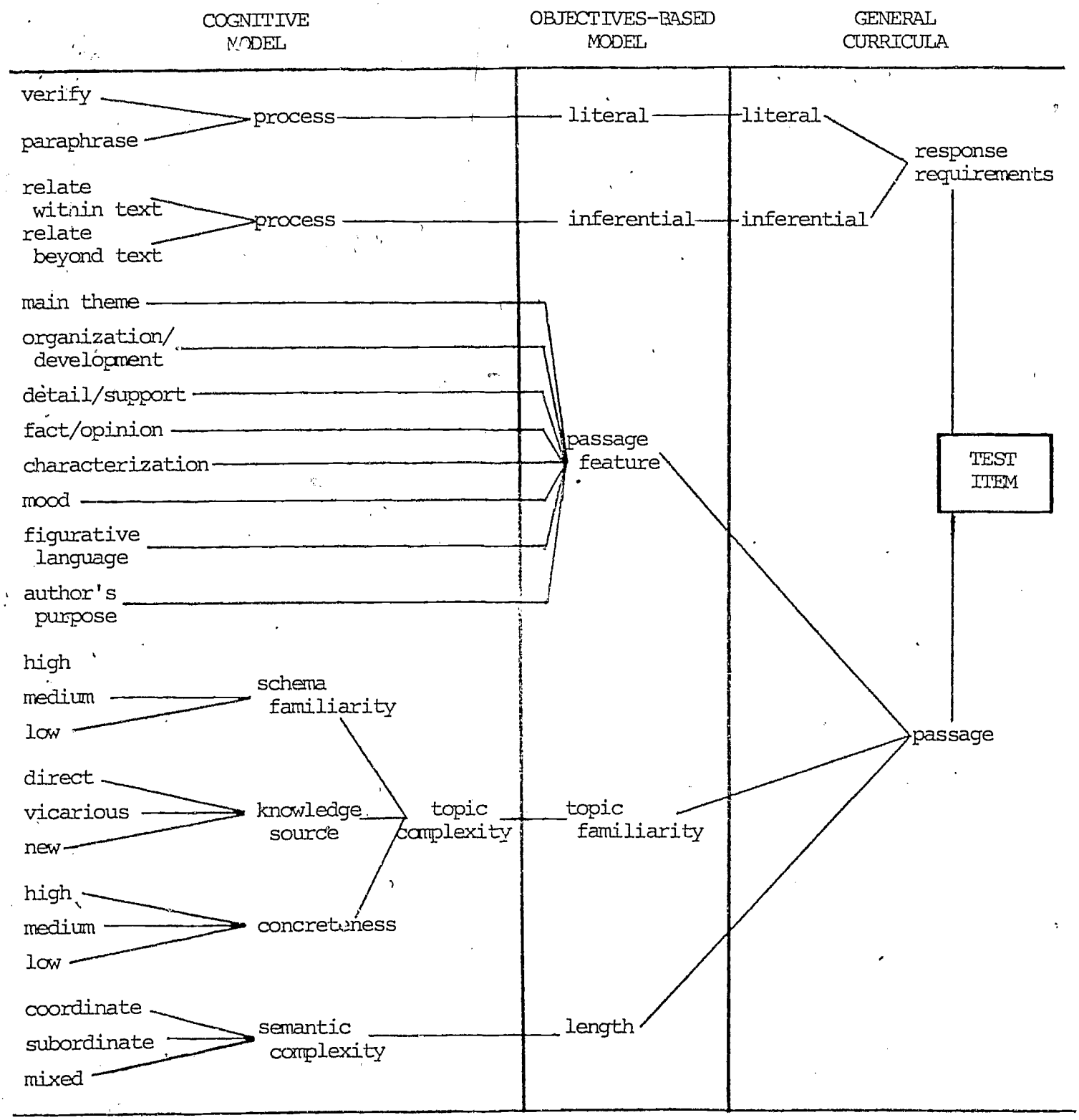

Figure 1. Tree diagram for three task structure schemes.

passage and those requiring a paraphrase of text information. Like an objectives-based scheme, a cognitive model examines performance on questions about text features such as main idea or fact and opinion. A cognitive model further examines performance on passages where topics differ according to students' probable familiarity with the general topic, the source of their knowledge (e.g., direct experience vs. reading) and the concreteness of the topic. A cognitive model might also compare students' performance on passages differing in semantic complexity, i.e., whether the relationship among ideas are primarily coordinate, subordinate or mixed. The issue in this phase of the study was whether the more detailed processing and passage dimensions proposed in a cognitive model of task structure would differentiate students' performance.

Procedure

Following the method used in Phase I of the study, two researchers used the cognitive model scheme to categorize the test items and reading passages. Appendix A contains the directions for scoring and definitions of the features. Appendix B records the groupings of the items and passages. Agreement between the raters on the categories ranged from .71 to 1.00 with an average of .92 (see Table 2). These detailed item groupings were then aggregated into the more global categories characteristic of objectives-based and general curricula schemes.

- - - - - - - - - -

Insert Table 2 here

- - - - - - - - - -

TABLE 2

Rater Agreement Levels

Task Feature	CRT	SAT
Skill	.96	.93
Task Genre	1.00	1.00
Text Feature	.96	.93
Response Complexity	.96	.99
Familiarity	.84	.76
Concreteness	.78	.93
Knowledge Source	.71	.90
Structural Complexity	1.00	.93
Length	1.00	1.00
TOTAL	.91	.93

RESULTS

Item classifications

Use of the scheme permits inspection of the distribution of the types of tasks presented within a test as well as comparison of types of tasks on the two tests. Table 3 presents the numbers and percentages of the items falling in each category.

- - - - - - - - - - -
Insert Table 3 here
- - - - - - - - - - -

On the norm-referenced test (NRT), the distribution of items requiring literal as opposed to inferential procedures is 39% vs. 61%. Forty-seven percent of the items relate to expository passages, forty-five to narrative, non-literary passages, and eight percent to a poem. Eighty-six percent of the text features questioned relate to details within the passage; fourteen percent ask for main ideas. Twenty-eight percent of the questions are on passages judged to be very familiar to sixth-graders; the rest of the questions are about topics judged as medium or low in familiarity. The texts are evenly distributed into high versus low or medium levels of concreteness. Approximately half of the questions are about content new to the students and another thirty-four percent are on topics students might have learned about in school, from TV, or by observation. Only twenty-five percent of the passages are structured using primarily coordinate relationships among ideas.

The district's criterion-referenced test (CRT) also asks more inferential (65%) than literal (35%) questions. Only 27% of the questions are

TABLE 3

Classification of Norm-referenced and Criterion-referenced Reading Comprehension Test Items According to Three Task Structure Schemes

Test	Cognitive Model Scheme				Objectives Based Scheme				General Curricula Scheme			
	NRT		CRT		NRT		CRT		NRT		CRT	
	%	f (71)	%	f (26)	%	f (71)	%	f (26)	%	f (71)	%	f (26)
Skill												
literal	39	(27)	35	(9)	32	(28)	35	(9)	39	(28)	35	(9)
inferential	61	(43)	65	(17)	61	(43)	65	(17)	61	(43)	65	(17)
Task Genre												
expository	47	(33)	27	(7)	--	--	--	--	--	--	--	--
persuasive		(0)		(0)	--	--	--	--	--	--	--	--
narrative literary	1	(1)	23	(6)	--	--	--	--	--	--	--	--
poem	7	(5)	0	(0)	--	--	--	--	--	--	--	--
non-literary	45	(32)	50	(13)	--	--	--	--	--	--	--	--
Text Feature												
main theme	14	(10)	12	(3)	14	(10)	12	(3)	--	--	--	--
organization/ development		(0)	8	(2)		(0)	8	(2)	--	--	--	--
detail/support	75	(53)	39	(10)	86	(61)	81	(21)	--	--	--	--
fact/opinion		(0)	8	(2)	--	--	--	--	--	--	--	--
characterization	6	(4)	12	(3)	--	--	--	--	--	--	--	--
mood		(0)	12	(3)	--	--	--	--	--	--	--	--
figurative language	6	(4)		(0)	--	--	--	--	--	--	--	--
author's purpose		(0)	12	(3)	--	--	--	--	--	--	--	--
Response Complexity												
Recognition	--	--	--	--	39	(28)	31	(8)	39	(28)	31	(8)
verify	8	(5)	8	(2)	--	--	--	--	--	--	--	--
paraphrase	31	(22)	23	(6)	--	--	--	--	--	--	--	--
Infer	--	--	--	--	61	(43)	65	(17)	61	(43)	65	(17)
infer relationship with text	28	(20)	32	(8)	--	--	--	--	--	--	--	--
infer relationship beyond text	33	(23)	38	(10)	--	--	--	--	--	--	--	--

TABLE 3 (Continued)

Test	Cognitive Model Scheme				Objectives Based Scheme				General Curricula Scheme			
	NRT		CRT		NRT		CRT		NRT		CRT	
	%	f (71)	%	f (26)	%	f (71)	%	f (26)	%	f (71)	%	f (26)
Passage Complexity												
topic/scheme												
familiarity												
high	28	(20)	15	(4)	--	--	--	--	--	--	--	--
medium	32	(28)	31	(8)	--	--	--	--	--	--	--	--
low	34	(24)	54	(14)	--	--	--	--	--	--	--	--
concreteness												
high	52	(37)	42	(11)	--	--	--	--	--	--	--	--
medium	24	(17)	27	(7)	--	--	--	--	--	--	--	--
low	24	(17)	31	(8)	--	--	--	--	--	--	--	--
knowledge source												
personal experience	12	(6)		(0)	--	--	--	--	--	--	--	--
vicarious/ school	34	(24)	54	(14)	--	--	--	--	--	--	--	--
new	54	(41)	46	(12)	--	--	--	--	--	--	--	--
Structural Complexity												
coordinate	25	(18)	73	(10)	--	--	--	--	--	--	--	--
subordinate	41	(29)		(0)	--	--	--	--	--	--	--	--
mixed	34	(24)	27	(7)	--	--	--	--	--	--	--	--
Length												
50-100	51	(36)	42	(11)	51	(36)	42	(11)	--	--	--	--
100-300												
300-500	49	(35)	58	(15)	49	(35)	58	(15)	--	--	--	--
Reading Level (1-13)	6		6		6		6		6		6	

in relation to expository passages; approximately half are about non-literary narrative text. Twenty-three percent (6) of the questions are about a literary fiction passage. The CRT asks two questions about the sequence of events and a few questions about characters (3) and about facts and opinions (2). Six of the eight literal questions require students to find a paraphrase of the correct answer in the text, and 10 of the 18 inference questions require students to draw upon knowledge not given in the text to make the inference. Questions relating to three of the six passages presented were judged to be on topics for which students were not likely to have much, if any background knowledge. The structure of the passages tended to present simpler, more coordinate relationships among ideas than did the structures of the passages on the NRT.

These descriptive data indicate that the two tests are not measuring reading comprehension in the same way. If we look at the proportions of literal and inferential questions, the tests look similar. However, if we look more closely at what students read and at what features of the texts they are asked to recognize or interpret, the two tests differ substantially. On the NRT, almost all the questions (92%) are in relation to expository or non-literary text, yet most of the material read by elementary students is in the literary narrative mode. About half the questions on the NRT are on expository text; only twenty-seven percent of the CRT test is on exposition. Cognitive research suggests that elementary students may have less well-formed schema for expository discourse structures and therefore these items would be more difficult than narrative structures. Furthermore, neither test places the same emphasis on narrative fiction that classroom

curricula do. Interestingly, the NRT presents a poem as its literary selection, while the CRT presents a story.

Approximately 60% of the questions on the NRT are in relation to reasonably familiar content; but only 48% of the CRT questions are about passages with familiar content. A low 18% of the questions from the NRT are on information students would have acquired through direct experience, and for which they would, presumably, have a well developed schema. None of the CRT passages tap such information. Many items on both tests, then, do not permit students to draw upon well established background knowledge about a topic to help them understand and interpret the test passage.

The tests differ most dramatically in the text features targeted by the questions. The tests ask a comparable proportion of main idea question (14% and 11%), but both percentages are low considering the emphasis placed on main idea in curricular goals. Fully seventy-five percent of the NRT questions ask for isolated details, in comparison to 39% detail questions on the CRT. Current research on the development of skilled reading performance criticizes standardized tests and materials that emphasize unconnected, often trivial details at the expense of questions that require students to identify details relevant for a broader gist or interpretation (Bauman, 1982; National Assessment of Educational Progress, 1981; Pearson, 1975).

Not only do the two tests differ in their emphasis on detail questions, but approximately 40% of the questions on the CRT are about text features such as author's purpose, mood, organization, and fact and opinion which are not queried by the standardized test at all.

In sum, the use of a cognitive model to classify test items according to the processes they require and the nature of the problem they present in the reading passage seems to provide a test profile that would be masked by a more global analysis. In this phase of the study and in Phase I the cognitive model task structure scheme revealed marked differences in how standardized, criterion-referenced and curriculum-embedded tests define and measure reading comprehension. The analyses of the content of the tests indicates that they are measuring different aspects of reading and with quite different degrees of emphasis. They do not measure the same thing --a finding consistent with other studies of reading tests and their relationsips to each other, instruction, and reading research (e.g., Bauman, 1982; Jenkins & Pany, 1976; Langer, 1981).

Response patterns for item clusters

Performance levels. In the first analyses of student responses to items in each cluster, the proportion of the 640 students who took both the tests and who answered an item correctly (p-value) was calculated. The p-values for individual items were then averaged for each cluster. In Table 4 the average p-values for each task structure category appears along with the range of p-values within each cell.

- - - - - - - - - - -
Insert Table 4 here
- - - - - - - - - - -

The most dramatic pattern in the table is the consistent difference in cluster difficulty between NRT and CRT items. The range of NRT p-values is only from .071 to .424, while the CRT p-values range from .677 to .893.

TABLE 4

Mean p Values*

	NRT	CRT
Skill		
literal	.264 (.001-.804)	.716 (.493-.903)
inferential	.192 (.002-.734)	.772 (.583-.931)
Task Genre		
expository	.265 (.042-.804)	.751 (.605-.856)
persuasive	.002	0
narrative literary	.177 (.059-.453)	.766 (.640-.882)
non-literary	.187 (.001-.734)	.748 (.493-.931)
Text Feature		
main theme	.182 (.010-.634)	.741 (.583-.910)
organization/development	0	.893 (.882-.903)
detail/support	.243 (.001-.804)	.708 (.533-.827)
fact/opinion	0	.724 (.716-.731)
characterization	.071 (.013-.208)	.706 (.493-.855)
mood	0	.805 (.640-.931)
figurative language	.138 (.077-.203)	0
author's purpose	0	.826 (.796-.856)
Response Complexity		
Recognition		
verify	.355 (.013-.775) [1]	.677 (.605-.749)
paraphrase	.255 (.001-.804) [1]	.702 (.493-.903)
Infer		
infer relationship within text	.263 (.010-.679) [1]	.804 (.730-.882)
infer relationship beyond text	.128** (.007-.428) [2]	.758 (.583-.931)

* Mean, with ranges in parentheses.

** Significantly less at $p<.05$ (order among means according to Newman-Kuels in brackets).

TABLE 4 (Continued)

	NRT	CRT
Passage Complexity		
topic/scheme		
familiarity		
high	.149 (.002-.618)	.762 (.659-.910)
medium	.201 (.001-.775)	.790 (.533-.931)
low	.297 (.042-.804)	.729 (.493-.882)
concreteness		
high	.215 (.001-.804)	.734 (.493-.882)
medium	.217 (.010-.679)	.740 (.605-.910)
low	.229 (.049-.734)	.790 (.533-.931)
knowledge source		
personal experience	.318 (.045-.679)	0
vicarious/school	.171 (.001-.775)	.774 (.553-.931)
new	.242 (.010-.804)	.728 (.493-.856)
Structural Complexity		
coordinate	.139 (.001-.679)	.754 (.493-.931)
subordinate	.272 (.002-.804)	0
mixed	.252 (.049-.775)	.751 (.605-.856)
Length		
50-100	.248 (.002-.804)	0
100-300	.199 (.045-.679)	.761 (.533-.931)
300-500	.213 (.001-.775)	.747 (.493-.882)

Yet both tests are presented as measures of sixth-grade reading comprehension.

The NRT items were, on the average, very difficult with p-values averaging around .250. The CRT on the other hand, was much easier with p-values of about .70. On both tests the performance range was quite restricted. The restrictions in range on both tests severely weakened the interpretability of analyses based on correlational techniques.

The average p-values in each of the task structure categories were examined to see if the levels of performance conformed to patterns that would be predicted by research and also to see if performance on CRT and NRT items in a particular category was comparable.

For items classified according to skill area and level, research suggests that performance on literal questions should exceed performance on inferential questions. This prediction is bourne out on the NRT, but not on the CRT. Within the further subdivisions of these skill areas into levels of response complexity, performance was highest (.424) on questions requiring verification in the text of information presented verbatim in the question and lowest (.128) on questions requiring students to draw relationships between passage content and information not given in the text. However, the difficulties of questions requiring paraphrase of text content are fairly comparable to questions asking students to draw inferences between ideas in the text. It may be that both types of questions require some inference. Even in paraphrasing, students must use schematic knowledge to recognize the relationship between words or sentences in the text and a synonym or transformation in a distractor. On the CRT, the verifica-

tion questions were harder than the paraphrase or inference questions, although these figures are based on only two verification questions.

For questions referenced to passages of distinct discourse structures, research might predict that items for literary narratives would be easiest since elementary students have more experience reading stories and therefore may have a more developed schemata for story structures. Next in difficulty would be non-literary narrative followed by exposition. This progression in difficulty from stories to exposition did not occur in the NRT and was only slightly apparent in the CRT. Therefore expectations for a dominant effect of discourse structure on comprehension were not corroborated in these tests.

Among the text features targeted by questions, the more global features of main theme, author's purpose and mood might have been expected to be harder. Such items require the combination of several pieces of text information and their integration with other background information. In the NRT, performance on main theme questions was very low (.187) while performance on details was greater, albeit low (.243). The CRT presents three items for each of three global categories, main theme, author's purpose and mood. The average p-values on these (.741, .826, .805) are higher than the average p-value for items asking for details (.708), although the small number or items requires interpreting these levels of performance with caution.

Two dimensions of passage complexity did not differentiate among levels of performance, judged familiarity and concreteness of the general topic. The judged source of student's information about a topic, i.e., direct experience and vicarious experience as opposed to totally new information did

corroborate research suggesting that students comprehend text better that is about topics on which they have more background information. On the NRT, items about passages for which it was judged that students might have learned the topic through direct personal experience were considerably easier (.318) (although performance was still low) than performance on items about passages on topics students may have learned about in school (.177) or on topics that seemed to be on totally new information (.242). On the CRT, there were no items for passages on topics students might have learned about through personal experience. However, the p-values for passages judged to present information that could have been vicariously learned or that was new follow the expected pattern.

Various methods of discourse analysis propose that the number of ideas or propositions within a text and the nature of their relationship to each other influence the comprehension of the passage. In the cognitive model task structure scheme, the measures of passage difficulty attributable to semantic structure of discourse were its structural complexity and its length. Research suggests that passages where ideas are coordinate are easier to comprehend than those presenting part/whole subordinate relationships or those presenting a mixture of the two. The categories of coordinate, subordinates and mixed revealed no pattern, suggesting that these categories may be too gross a classification of passage structure. Performance on passages of different lengths and therefore, number of propositions was slightly higher on shorter passages.

The most striking findings of the preceding analyses are the large differences in difficulty levels between the NRT and CRT and the restricted

range of scores in each. The average difficulty levels of items within the various task structure categories supported only some of the expectations based on theories of learning and reading development, although the small numbers of items in some of the categories, particularly on the CRT, require drawing any conclusions with caution.

No single dimension seemed to have a dominant influence on performance levels. This finding may be reasonable since research on any one factor such as background knowledge or discourse structure tends to equate problems or passages on the other dimensions. In an attempt to examine the homogeneity of items with more than one task structure in common, the items were grouped according to their match on the features of general skill level, task genre, text features, response complexity and knowledge source. Table 5 presents these item classification patterns for the NRT and CRT simultaneously.

- - - - - - - - - - -

Insert Table 5 here

- - - - - - - - - - -

These tables depict even more dramatically than Table 2 the lack of homogeneity of items at either the descriptive or performance levels. For example, on the NRT, there are eight items requiring literal comprehension in which the answer is a paraphrase of a detail in an expository passage on vicariously learned information. Performance on these items ranged from .002 to .534. On the CRT there were almost no items that matched on all five dimensions preventing any more precise comparisons.

Item analyses. The second set of analyses tested the homogeneity of items within the task clusters and the relationship between NRT and CRT

TABLE 5

ITEM CLASSIFICATION PATTERNS

Test		Classifications					$\bar{X}$ p Values[6]
NRT	CRT	Skill[1]	Task Genre[2]	Text Feature[3]	Response Complexity[4]	Knowledge Source[5]	
x		1	1	3	2	2	.305 (8)
x		1	1	3	2	3	.329 (6)
x		1	1	3	1	2	.775 (1)
	x	1	1	3	1	3	.605 (1)
x		1	1	3	4	3	.087 (2)
	x	1	3	2	3	2	.882 (1)
	x	1	3	3	2	2	.646 (1)
x		1	3	3	1	3	.059 (1)
	x	1	4	2	2	2	.903 (1)
	x	1	4	3	2	2	.608 (2)
x		1	4	3	2	2	.030 (3)
	x	1	4	3	1	2	.931 (1)
	x	1	4	3	2	3	.810 (1)
x		1	4	3	2	3	.024 (1)
x		1	4	3	1	3	.324 (4)
x		1	4	5	2	2	.033 (1)
	x	1	4	5	2	3	.493 (1)
	x	2	1	1	3	3	.730 (1)
x		2	1	1	3	3	.400 (2)
x		2	1	1	3	2	.120 (2)
x		2	1	1	4	2	.207 (1)
x		2	1	3	3	3	.289 (2)

[1]Skill: (1) literal, and (2) inferential.

[2]Task Genre: (1) expository, (2) persuasive, (3) narrative-literary, and (4) narrative-non-literary.

[3]Text Feature: (1) main theme, (2) organizational, (3) details, (4) fact/opinion, (5) characterization, (6) mood, (7) figurative language, and (8) author's purpose.

[4]Response Complexity: (1) verification, (2) paraphrase, (3) inferred from within text, and (4) inferred from beyond text.

[5]Knowledge Source: (1) personal experience, (2) vicarious or school, and (3) new.

[6]Means, with number of contributing items in parentheses.

TABLE 5 (Continued)

ITEM CLASSIFICATION PATTERNS

Test		Classifications					$\bar{X}$ p Values[6]
NRT	CRT	Skill	Task Genre[2]	Text Feature[3]	Response Complexity[4]	Knowledge Source[5]	
	x	2	1	3	3	3	.777 (2)
x		2	1	3	3	2	.160 (2)
x		2	1	3	4	2	.193 (3)
x		2	1	3	4	3	.101 (3)
	x	2	1	4	4	3	.716 (1)
x		2	1	7	4	3	.077 (1)
	x	2	1	8	3	3	.856 (1)
	x	2	1	8	4	3	.796 (1)
x		2	3	1	3	3	.453 (1)
	x	2	3	3	3	2	.801 (1)
x		2	3	3	4	3	.073 (1)
	x	2	3	5	3	2	.855 (1)
	x	2	3	5	4	2	.771 (1)
	x	2	3	6	4	2	.640 (1)
x		2	3	7	3	3	.085 (1)
x		2	3	7	4	3	.194 (2)
x		2	4	1	3	2	.017 (2)
x		2	4	1	4	1	.045 (1)
x		2	4	1	4	2	.045 (1)
	x	2	4	1	4	2	.910 (1)
	x	2	4	1	4	3	.583 (1)
x		2	4	3	2	3	.326 (4)
x		2	4	3	3	1	.627 (2)
x		2	4	3	3	2	.041 (1)
x		2	4	3	3	3	.408 (1)
	x	2	4	3	3	3	.754 (1)
x		2	4	3	4	1	.203 (3)
x		2	4	3	4	2	.099 (4)
x		2	4	3	4	3	.157 (1)
	x	2	4	4	4	2	.731 (1)
x		2	4	5	4	2	.084 (3)
	x	2	4	6	4	3	.843 (1)
	x	2	4	7	4	2	.659 (1)

items in the same clusters by comparing the regularity of responses to items within the clusters. The analysis used was a system developed in Japan by Sato and his colleagues which examines patterns of student responses on a test (Sato, 1974). The S-P technique arrays test scores in a Student-Problem matrix in which rows represent individual responses and columns represent group responses to the set of items. Rows are ordered by descending total number of correct responses and columns are ordered by ascending order of item difficulties. The procedure then measures the degree to which the cumulative ogive curves of the student performance (S-curve) and problem difficulties (P-curve) overlap. Perfect overlap would look much like a Guttman scale. As the pattern of responses becomes increasingly random, the curves become more discrepant. Sato had developed an index, termed the Caution index, (C), which calculates the degree of discrepency between the curves. Perfectly matched curves produce a Caution index of 0, a completely random pattern will approach 1.0. Therefore, a high index value for a respondent or item signals that performance is discrepant from the pattern established by all members of the set. (See McArthur, 1982, (a), (b). In practice, a Caution index above .30 has been considered a signal of an aberrant pattern. Because of its simplicity, the S-P procedure has been used extensively in Japan for the analyses of tests, items, and instructional hierarchies and for feedback to students and teachers. Table 6 displays the average Caution index and range of Caution indices for each set of NRT and CRT items clustered according to the cognitive task structure scheme.

- - - - - - - - - -
Insert Table 6 here
- - - - - - - - - -

TABLE 6

Caution Indices from SATO Analysis*

	NRT		CRT	
Skill				
literal	.401	(.265-.667)	.214	(.143-.294)
inferential	.354**	(.256-.493)	.251	(.147-.363)
Task Genre				
expository	.383	(.256-.464) [1]	.245	(.216-.343) [1]
persuasive	0		0	
narrative literary	.299**	(.279-.327) [2]	.196**	(.143-.256) [2]
non-literary	.369	(.256-.667) [1]	.238	(.175-.363) [1]
Text Feature				
main theme	.347	(.260-.415)	.313	(.267-.363) [1]
organization/development	0		.166**	(.143-.188) [3]
detail/support	.379	(.256-.667)	.256**	(.175-.343) [2]
fact/opinion	0		.248**	(.233-.263) [2]
characterization	.388	(.343-.428)	.197**	(.157-.233) [3]
mood	0		.237**	(.224-.252) [2]
figurative language	.309	(.279-.364)	0	
author's purpose			.233**	(.216-.249) [2]
Response Complexity				
Recognition				
verify	.375	(.297-.436) [1]	.224	(.204-.243)
paraphrase	.402	(.275-.667) [1]	.223	(.177-.294)
Infer				
infer relationship within text	.351**	(.256-.449) [2]	.228	(.143-.343)
infer relationship beyond text	.355**	(.260-.464) [2]	.258	(.216-.363)

* Means, with ranges in parentheses.

** Significantly less at $p<.05$.

[] Order among means by Newman-Kuels ($p<.05$).

TABLE 6 (Continued)

	NRT		CRT	
Passage/Complexity				
topic/scheme				
familiarity				
high	.395	(.295-.493)	.269	(.204-.363) [1]
medium	.359	(.256-.667)	.268	(.188-.343) [1]
low	.366	(.279-.464)	.213**	(.143-.308) [2]
concreteness				
high	.410	(.340-.667) [1]	.201**	(.143-.267) [2]
medium	.326**	(.256-.449) [2]	.263	(.204-.363) [1]
low	.335**	(.279-.394) [2]	.268	(.188-.343) [1]
knowledge source				
personal experience	.300**	(.256-.336) [2]	0	
vicarious/school	.388	(.256-.667) [1]	.231	(.143-.363)
new	.369	(.279-.464) [1]	.246	(.175-.343)
Structural Complexity				
coordinate	.373	(.256-.667)	.225**	(.143-.363)
subordinate	.352	(.256-.493)	0	
mixed	.387	(.299-.464)	.275	(.216-.343)
Length				
50-100	.407	(.343-.493) [1]	0	
100-300	.322**	(.256-.393) [3]	.257	(.188-.363)
300-500	.378**	(.256-.667) [2]	.224	(.143-.343)

In general, the Caution indices for the NRT items are above .30, suggesting that the pattern of responses on the NRT does not follow any discernably regular pattern. This finding is surprising since norm-referenced tests are presumably constructed to present items that discriminate systematically between high and low scoring examinees. The CRT Caution indices are lower than the indices for the NRT items and this difference is significant. ($F(1,95)=32.01$, $p<.001$)

Inspection of performance on items within each task structure category reveals first that, on the NRT, the average Caution index for literal comprehension questions (and the range) is significantly higher than the average Caution index for inferential questions ($F(1,70)=5.35$, $p<.05$). Concommitantly, the average Caution indices for items requiring verification or paraphrasing procedures are significantly higher (.375, .402) than the average indices for items requiring inferences within or beyond the text (.228, .258; $F(3,68)=4.08$, $p<.05$). These differences may imply that the sorts of information and fact-finding skills measured by the large number of detail questions on the NRT are less homogeneous than the skill labels imply. On the CRT, on the other hand, the average Caution indices for literal and inferential items and for their subdivisions into the response complexity categories are low and not significantly different from each other.

In the Task Genre categories, the NRT items related to the poem have a lower average Caution index, a narrower range of indices, and, therefore, a more predictable pattern of performance than do items in the other Genre categories. This information, coupled with the p-values in Table 3, suggests

that the items related to the poem were consistently harder and more discriminating than items on passages with other discourse structures. On the CRT, for items grouped according to the genre of the reading passage, performance on items related to the literary narrative passage (a story) was significantly more symmetrical than performance on items referencing expository or non-literary passages ($F(2,23)=4.81$, $p<.05$).

None of the Caution indices for Text Feature categories on the NRT were significantly different from each other, although they were all high. On the CRT, the average Caution index for the three items asking main theme questions was significantly higher than the average Caution indices for the other text features. ($F(7,63)=3.02$, $p<.05$) This result may imply that the information and strategies necessary to derive the main idea of the three passages questioned were highly diverse and, therefore, these were not homogeneous items.

The Caution indices for passage complexity further describe how dimensions of the reading text affected performance on items related to the passage. The NRT Caution indices for topic/schema familiarity were higher than those for the CRT, but the NRT indices did not differ significantly from each other.

CRT items about passages with topics judged to be least familiar to students yielded Caution indices significantly lower than the indices for medium and high familiarity passages. ($F(2,23)=5.34$, $p<.05$) P-values for these items were also considerably lower.

The category of concreteness also produced performance patterns contrary to those on the CRT. NRT items for passages presenting medium and

and low concrete content yielded Caution indices significantly lower than the Caution indices for highly concrete passages. Conversely, on the CRT items about the two passages judged to present highly concrete material produced a Caution indices significantly lower than those for the other two levels of concreteness.

Performance patterns on items grouped according to the judged sources of students' information about the content are more regular and interpretable. On the NRT, the average Caution index for passages and range of indices with content referencing personal experience is significantly lower (and borderline acceptable) than patterns on the other two categories ($F(2,69)=4.35$, $p<.05$). The influence of knowledge source on CRT performance patterns follows the same pattern but is not significant.

Structural complexity does not significantly differentiate among performance patterns on NRT items, but for CRT items about the four passages presenting primarily coordinate information, the Caution indices are significantly lower and in a narrower range than passages presenting mixed idea structures. On the NRT, patterns of performance are variable according the length of the reading passage, ($F(2.69)=3.81$, $P<.05$) but length does not differentiate among CRT item performance patterns.

According to the Caution indices, performance patterns on the NRT are so erratic that the items are highly suspect. Of the seventy-one items, 58 have Caution indices over .3. In contrast, CRT patterns of performance tend to be more stable and predictable. On the CRT, 3 of the items have Caution indices over .30. These data cast into doubt the technical adequacy of the NRT for this population.

The ranges of the p-value and Caution indices can jointly provide some indication of the homogeneity of items in the cognitive task structure categories. On the NRT, p-values range from .001-.804, and Caution indices range from .256-.667. The range of CRT p-values is .493-.931, the range of Caution indices is .143-.363. Performance within many categories of the cognitive task structure scheme yielded narrower ranges of p-values and Caution indices, implying that items in these categories may be more homogeneous than items in the less stable categories.

Factor Analyses

In a final set of analyses, both tests were subject to standard factor analysis procedures using SPSS. Oblique rotation loading matrices for both tests resulted in meaningless loading structures which only reflected item difficulty groupings. In the case of the NRT data, seven factors emerged, each representing items clustered according to common difficulty levels (p values) but which had no discernable pattern of item content in common. For CRT data, only two factors emerged, the first accounting for 24 of the 26 items. As with NRT data, the latter factors again reflected shared p value ranges.

SPSS factor analyses are based on manipulations of Pearson correlation coefficients. In the case of dichotomous (correct versus incorrect) data, Pearson coefficients (Phi coefficients in the dichotomous case) are highly affected and distorted by item difficulty thresholds. What should provide a more clearly interpretable outcome would be analogous factor analyses functions based on a coefficient of correlation which is not so adversely

affected by the difficulty of test items. Such a coefficient is found in tetrachoric correlation coefficients.

Bengt Muthén of UCLA has developed an exploratory factor analysis model based on tetrachoric correlations. Both tests were reanalyzed using Muthén's program. The NRT analysis revealed a factor solution ($X^2_{(943)}$=1005.425), of which the first two factors had significant eigen values of 5.33 and 1.83 respectively. Items loading on the factors reflected an interesting structure, which was most easily interpreted as a principally "literal" first factor and an "inferential" second factor. These labels reflected classification of items through both Skill and Response Complexity categories. The third factor (eigenvalue <.600) did not represent any consistent pattern for item loadings.

- - - - - - - - - - -

Insert Table 7 here

- - - - - - - - - - -

The analysis of CRT data produced a two factor solution ($X^2_{(298)}$=253.77). While NRT factors appeared to represent levels of the Skill category classifications, CRT factors were more associated with Familiarity categories. Factor I was a "Low Familiarity" cluster, while items on Factor II were "High and Medium Familiarity."

The conclusion of interest in these analysis procedures is that the use of traditional, more global analysis techniques does appear to support claims for the simpler, grosser classification systems used in objectives- or curriculum-based models. However, the factor structures do <u>not</u> reflect the anticipated clusterings based on those simpler models. To the contrary, they mirror far more detailed views of what skills items appear to be reflect-

TABLE 7

Factor Analysis of the Reading Comprehension Tests

Factor Loadings*

	NRT		
Item	Factor I	Factor II	Factor III
17	.606	-	-
23	.324	-	-
24	.308	-	-
26	.608	-	-
27	.339	-	-
28	.374	-	-
30	.335	-	-
31	.299	-	-
33	.386	-	-
34	.642	-	-
35	.589	-	-
36	.371	-	-
39	.533	-	-
45	-	-	.310
46	-	-	.426
47	-	-	.296
48	-	-	.380
50	-	-	.323
51	-	-	.417
53	-	-	.403
54	-	-	.375
56	-	-	.331
57	-	-	.296
58	-	.575	-
59	-	.474	-
60	-	.625	-
63	-	.364	-
64	-	.575	-
67	-	.333	-
68	-	.684	-
70	-	.410	-
71	-	.470	-

*Only items with loadings >.30 are reported.

TABLE 7 (Continued)

Factor Analysis of the Reading Comprehension Tests

Factor Loadings*

	CRT	
Item	Factor I	Factor II
6	-	.791
7	-	.542
8	-	.465
9	-	.866
21	-	.455
22	-	.386
23	-	.360
27	.328	.405
28	.367	-
29	-	.598
30	.545	-
31	.493	-
43	-	.337
45	-	.388
46	-	.398
59	.557	-
60	.839	-
61	.640	-
62	.325	-
63	.721	-
64	.848	-
65	.300	-
66	.299	-
67	.410	-

*Only items with loadings >.30 are reported.

ing. Also noteworthy is the fact that the more traditional use of Phi coefficients may have led researchers in test development to erroneously rely on factors based purely upon item or task difficulty, not content.

Discussion

The purpose of the study was to explore the potential of a cognitive model task analysis scheme that specifies features of test problems shown by research to affect performance. The study examined the results of using the scheme to classify and compare the content of tests and then to analyze performance patterns of items clustered according to the task features.

Descriptive analyses. Use of the cognitive task structure scheme to analyze two reading tests reveals that they are presenting quite different types of passages and items to measure reading comprehension. The cognitive model of important variables in a task structure scheme presents features of the problem situation, (e.g., the reading passage) and the level of processing required to solve the problem (e.g., levels of literal and inferential discourse processing) that have a research base documenting their affect on performance. Application of the cognitive task structure scheme was reliable and revealed differences in the nature and distribution of items and passages. Analyses of the three tests analyzed in Phase I of the study also revealed large variations in test content.

Such a detailed picture of the structure of tasks presented on tests could be useful for mapping the degree of match between different tests, between tests and curricula and between tests and factors identified by research as important for distinguishing among levels of competence. More

precise specification of the range of concepts, strategies and problem formats actually covered by a test could help test designers produce more homogeneous pools of passages and test items. In this analysis, for example, there were relatively few clusters of items on the NRT and CRT that shared more than one dimension. This lack of comparability between test items purportedly measuring the same underlying construct weakens the psychological and practical rationales for comparing students' scores on different tests. It also limits the confidence with which users can generalize about the meaning of test scores.

Response patterns. Analyses of student responses to items in each cluster examined levels of performance and their distributions. The range and average proportion of correct responses for items in each cluster was calculated for the sixth-grade norm-referenced and criterion-referenced reading test. The substantial difference between performance on the NRT and CRT raises serious questions about the relative suitability of the tests for describing and interpreting performance. The high performance on the CRT portrays a picture of students' reading competencies quite different from the picture presented on the NRT.

One explanation for the highly discrepant difficulty levels of the items on the two tests comes from the descriptive analyses of the tests' content. They present different kinds of reading material and questions. Forty percent of the CRT items asked about text features not tested by the NRT at all. However, performance on questions about the text features querried on both tests is still highly discrepant.

The response pattern analyses using the S-P procedure provide another clue to the performance level discrepancies. Patterns of performance on the NRT items are highly erratic (above .30) on most of the items. According to the Sato procedure, these high Caution indices signal that many of the NRT items are not eliciting interpretable performance and requires further scrutiny. It may be that the test seemed so difficult to students that they resorted to unsystematic guessing. Even for a norm-referenced test, the dispersion of scores for this population is alarmingly unsystematic and requires more information about factors such as test administration conditions and the relationship of the structure of the test's passages and items to those students receive in instruction. Analyses of other NRT and CRT test contents and scores would be necessary to clarify the potential of the cognitive task analysis scheme.

Despite the limitations of the performance data available for this exploratory study, patterns of responses on the CRT and NRT imply that the cognitive task structure scheme may be a promising tool for describing, analyzing, and interpreting test performance. Score reports describing test scores according to the scheme could help teachers pinpoint not only the general skills, but the kinds of reading materials on which students need help, or are proficient. Performance patterns on many of the task structure dimensions were interpretable in light of cognitive research. In contrast to the ambiguous results of the more conventional factor analysis techniques for inducing skill constructs, the detailed partitioning of skill requirements used in the cognitive-based model seems to present a

clearer and more interpretable test analysis tool. Statistics used to profile performance can be simple (p-values and Caution indices) and easily understood by test users. Therefore, the cognitive task structure scheme and statistical techniques used in the study seem to offer more logical, comprehensible and psychologically sound methods for planning and interpreting what tests are testing.

APPENDIX A

Task Structure Schemes

Reading Comprehension

Directions: Read the passage and classify it for passage complexity.

Topic

Familiarity - general schema

students at the grade level are likely to have a well developed conception of the defining attributes of the passage's general topic (e.g., "bicycles" vs. "how to equip a 10-speed bike.")

Concreteness of passage-specific information

high - the passage describes events or details that the student can directly experience (touch, feel, see).

medium - the passage describes events or details in sensory terms, but the student is unlikely to be able to directly experience them. Includes historical accounts.

low - the passage describes abstract ideas (truth, love), unobservable phenomena (osmosis), or uses imprecise, abstract language.

Source - likely source of specific passage information

personal experience

vicarious experience - TV, movies, oral stories, books, school material, observed home and community events

new - the particular information is likely to be new to most American students at the age level.

Structural Complexity

coordinate - usually a series of events or set of ideas at the same level of generality

subordinate - general ideas and more specific details

mixed - some of both

Length

number of words in the passage

Reading level

as designated by the test/text publishers

APPENDIX A (continued)

Task Structure Schemes

Context

time - unlimited or limited

relevance

purpose for reading - writer's role specified

audience/function - use of information specified

General Skill

literal - the correct answer can be found in the passage, verbatim or as a paraphrase

inferential - the answer requires using information not available in the literal text, knowing the definition of a concept and its defining attributes in order to pick out an appropriate label for or instance given in the passage or an instance for a concept given in the passage (facts, opinions)

e.g., 1:

e.g., 2: knowing the usual sequence of events in a process or cause effect sequence (inferring motives based on knowledge of usual cause of actions)

Task Genre

exposition - a passage designed to inform

persuasive - presents an argument and reasons to move the audience to action or agreement

narrative -

literary - fictional stories with plots, characters, settings

non-literary - non-fiction account, e.g., of a series of events over time

Text Feature

the feature of the passage referenced in the question
main theme - the general point of issue or message about life
organization/development - questions about sequence, plot development
detail/support - questions about isolated facts, reasons, events -- often "wh" questions
fact/opinion - questions requiring the discrimination of factural statements versus statements of opinion, based on the text

APPENDIX B

PART 1

Classification of Text Items According to
A Cognitive Model of Task Structure:
Stanford Achievement Test Reading Comprehension (NRT)

Skill	
literal	1-3, 6, 8, 10, 14, 15, 18-21, 23, 26, 27, 29, 30, 33-35, 39, 44, 50, 52, 54, 62, 68
inferential	4, 5, 7, 9, 11-13, 16, 17, 22, 24, 25, 28, 31, 32, 36-38, 40-43, 45-49, 53, 55-61, 63-67, 69-71
Task Genre	
expository	1-6, 26-31, 32-36, 37-42, 50-54, 67-71
persuasive	
narrative (Poem) literary	61-66
non-literary	7-13, 14-19, 20-25, 43-49, 55-60
Text Feature	
main theme	4, 12, 16, 22, 31, 36, 37, 55, 64, 71
organization/ development	
detail/support	1, 2, 3, 5, 6, 10, 12-21, 23-30, 32-35, 38-52, 54, 56-60, 62, 66-70
fact/opinion	
characterization	7, 8, 9, 11
mood	
figurative language	53, 61, 63, 65
author's purpose	
Response Complexity	
Recognition - verify	15, 18, 26, 44, 62
paraphrase	1-3, 6, 10, 19, 20, 27, 29, 30, 33-35, 39, 46-49, 51, 52, 54, 68
Infer	
infer relationship within text	4, 12, 16, 28, 31, 36-38, 41, 45, 58, 60, 64, 65, 70
infer relationship beyond text	5, 7, 9, 11, 13, 14, 17, 24, 25, 32, 40, 42, 43, 53, 55-57, 59, 61, 63, 66, 67, 69, 71

APPENDIX B

PART 1 (CONTINUED)

Passage Complexity	
topic/scheme	
familiarity	
high	1-6, 7-13, 14-19
medium	20-25, 26-31, 50-54, 55-60, 67-71
low	32-36, 37-42, 43-49, 61-66
conreteness	
high	1-6, 7-13, 20-25, 26-31, 32-36, 37-42
medium	55-60, 14-19, 67-71
low	43-49, 50-54, 61-66
knowledge source	
personal experience	55-60
vicarious/school	1-6, 7-13, 20-25, 26-31, 67-71
new	14-19, 32-36, 37-42, 43-49, 61-66, 50-54
Structural Complexity	
coordinate	7-13, 14-19, 20-25, 55-60
subordinate	1-6, 32-36, 61-66, 67-71
mixed	26-31, 37-42, 43-49, 50-54
Length	
50-100	1-6, 7-13, 32-36
100-300	50-54, 55-60, 61-66
300-500	14-19, 20-25, 26-31, 27-71, 37-42, 43-49
Reading Level (1-13)	6

APPENDIX B

PART 2

Classification of Test Items According to
A Cognitive Model of Task Structure:
District Criterion Referenced Reading Test

Skill	
literal	6, 8, 9, 20, 28, 31, 63, 64, 66
inferential	7, 21, 22, 23, 27, 29, 30, 43-46, 59-62, 65, 67
Task Genre	
expository	43-46, 65-67
persuasive	
narrative (Story) literary	59-64
non-literary	6-9, 20-23, 27-31
Text Feature	
main theme	21, 29, 65
organization/ development	9, 64
detail/support	6, 8, 20, 21, 30, 44-46, 63, 66
fact/opinion	22, 43
characterization	28, 60, 61
mood	7, 27, 62
figurative language	
author's purpose	23, 45, 67
Response Complexity	
Recognition	
--verify	20, 66
paraphrase	6, 8, 9, 28, 31, 63
Infer	
infer relationship within text	30, 44-46, 59, 60, 64, 65
infer relationship beyond text	7, 21-23, 27, 29, 43, 61, 62, 67

APPENDIX B

PART 2 (CONTINUED)

Passage Complexity	
topic/scheme	
familiarity	
high	20-23
medium	6-9, 43-46
low	27-31, 59-64, 65-67
conreteness	
high	27-31, 59-64
medium	20-23, 65-67
low	6-9, 43-46
knowledge source	
personal experience	
vicarious/school	6-9, 20-23, 59-64
new	27-31, 65-67, 43-46
Structural Complexity	
coordinate	6-9, 20-23, 27-31, 59-64
subordinate	
mixed	65-67, 43-46
Length	
50-100	
100-300	6-9, 20-23, 65-67
300-500	27-31, 43-46, 59-64
Reading Level	
(1-13)	6

REFERENCES

Baker, E. L. Beyond objectives - Domain-referenced tests for evaluation and instructional improvement. Educational Technology, 1974, 14, 10-21.

Bauman, J. Linguistic structure and the validity of reading comprehension tests. Center for Applied Linguistics, Washington, D. C. Final report to the National Institute of Education (NIE-G-80-0149), May, 1982.

Glaser, R. The future of testing: A research agenda for cognitive psychology and psychometrics. American Psychologist, September, 1981, 36, No. 9.

Hambleton, R. K., & Simon, R. Steps for constructing criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

Hively, W. Introduction to domain-referenced testing. Educational Technology, 1974, 14, 15-20.

Jenkins, J. R., & Pany, D. Curriculum biases in reading achievement tests. (Technical Report No. 16) Urbana, IL: University of Illinois, Center for the Study of Reading, November 1976.

Langer, J. A. The construction of meaning and the assessment of comprehension: An analysis of reader performance on standardized test items. University of California, Berkeley, 1981. Under Review.

McArthur, D. L. Analyses of patterns - The S-P technique. In B. Choppin (Ed.) A critical comparison of psychometric models for measuring achievement. NIE Final Report (NIE-G-80-0112, P3). Center for the Study of Evaluation, Graduate School of Education, University of California, Los Angeles, 1982. (a)

McArthur, D. L. Detection of item bias using analyses of response patterns. Paper presented at the annual meeting of the American Educational Research Association, New York, 1982. (b)

National Assessment of Educational Progress. Reading, thinking, and writing: Results from the 1979-80 National Assessment of Reading and Literature. Denver, Colorado, 1981.

Pearson, D. D. The effect of grammatical complexity on children's comprehension, recall, and conception of certain semantic relations. Reading Research Quarterly, 1975, 10, 155-192.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, N.J.: Prentice Hall, 1978.

Quellmalz, E. S. Implications of learning research for test design. In Baker, E., Linn, R., & Quellmalz, E. Knowledge synthesis of criterion-referenced testing. Final report to the National Institute of Education (NIE-G-79). Under review. (1981)

Sato, T. A classroom information system for teachers, with focus on the instructional data collection and analysis. Association for Computer Machinery Proceedings, 1974, 199-206.

Sternberg, R. J. Testing and cognitive psychology. American Psychologist, October 1981, 36, No. 10, 1181-1189.