4548

ASSESSING THE CONTENT STANDARDS OF A LARGE-SCALE, STANDARDS-BASED TEST: A PSYCHOMETRIC VALIDITY STUDY OF THE 2002 HAWAIʻI STATE ASSESSMENT GRADE 8 AND GRADE 10 READING TESTS

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE
UNIVERSITY OF HAWAIʻI IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

EDUCATIONAL PSYCHOLOGY

DECEMBER 2004

By
Russell K. Uyeno

Dissertation Committee:

Shuqiang Zhang, Chairperson
Ann Bayer
Selvin Chin-Chance
Lois Yamauchi
Daniel Spears

ACKNOWLEDGEMENTS

I would like to thank my classmates in the EDEP 768 course focused on the Hawai‘i State Assessment that met during the fall semester of 2003, taught by Dr. Shuqiang Zhang, for providing ideas and feedback that were most valuable in framing this study. Particular thanks go to Mr. Gerald Plett, an employee of the Hawai‘i State Department of Education (DOE) and member of the class, and to the DOE Planning and Evaluation Office, for providing invaluable assistance to me and other group members by making the test data available in a convenient and user-friendly format, and for responding to requests for clarification of the data.

ABSTRACT

The widespread use of large-scale, standards-based testing to measure educational achievement and, by extension, instructional effectiveness, administrative efficiency, and progress towards policy goals, has created a clear need for ongoing empirical study to validate the appropriate uses of the resulting test scores. An important part of these validation efforts concerns the relationship between the content standards that underlie the structure and content of a test, and the scores obtained in actual test administrations.

Analyses were conducted on the Grade 8 and Grade 10 reading scores from the 2002 Hawaiʻi State Assessment (HSA) test. These analyses examined relationships among the scores in the following areas: (a) correlations between norm-referenced items specifically designated as being aligned with the HSA content standards, and standards-based items based on the HSA standards; (b) correlation patterns among items aligned with the three components of reading ability embodied in the HSA test; (c) the uniqueness of information provided by constructed-response items, over and above that provided by multiple-choice items; and (d) the relative difficulty of passage types in relation to the student's overall reading proficiency level.

The results did not support the distinction between aligned and nonaligned norm-referenced items; both aligned and nonaligned items appear to measure the same construct that the standards-based items measure. Also, the results did not support the three-component model of reading ability that underlies the structure of the test, thereby calling into question the use of component subscores to indicate areas of relative strength and weakness in a student's reading ability, particularly in light of generally accepted standards for test interpretation and use. There was support for the uniqueness of information obtained from constructed-response items, although the magnitude of that information is small relative to that obtained by multiple-choice items. Finally, there was evidence that the difficulty of specific passage types depends on a student's overall reading proficiency.

The results of this study suggest that we may need to be more conservative in our expectations of the kinds of assessment information we can validly obtain from large-scale, standards-based reading tests. This conclusion is at odds with the premise that the content standards that underlie standards-based tests will result in discrete standard-specific assessment. In fact, the results of this study indicate that the discriminant validity, and the diagnostic value, of large-scale, standards-based assessment, cannot be psychometrically supported.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION TO THE STUDY

Standardized, large-scale testing has become a key feature of educational assessment in every state (Tindal, 2002). The idea that such tests should be based on, and aligned with, clear standards of what students should know and how well they should perform, has been widely embraced by states and educational jurisdictions, educational organizations, and other stakeholders as the most promising way to improve not only educational assessment but the educational system in general (Rothman, Slattery, Vranek, & Resnick, 2002). It has been noted that the educational reform movements based on the adoption of standards "has increased the amount of testing in K-12 schools and raised the consequences, expectations, and controversies attached to test results" (Pellegrino, Chudowsky, Glaser, & National Research Council (U.S.). Division of Behavioral and Social Sciences and Education. Committee on the Foundations of Assessment., 2001, p. 24). From the standpoint of assessment, one of the key advantages cited by proponents of standards-based testing (SBT) is its amenability to the incorporation of locally relevant curriculum content. Hence, SBT has the ability to deliver scores that are more meaningful to students, teachers, and schools than the information provided by the general measures of performance relative to national norms that traditional, norm-referenced testing (NRT) is designed to deliver. Properly constructed and interpreted, SBT thus holds much promise for improving the quality of assessment information for all those concerned with school education (Linn, 2002a).

It has been pointed out, however, that the rapid adoption of SBT has proceeded without a commensurate base of empirically grounded validity research. As Kifer (2001) observes, “It is, of course, difficult to argue against standards and even more difficult to be against high standards. . . .Yet, there are serious technical, conceptual, and pragmatic issues surrounding standards-based assessments” (p. ix). For SBT, there are two general standards-related areas that are of concern to validity investigations: content standards, which provide the framework for understanding the constructs and abilities that the test is designed to test, and performance standards, which provide the framework for understanding the levels at which the student has performed on those constructs and abilities, and on the test in general. Performance standards, which ultimately divide students into categories such as “meeting standards” and “below standards,” have attracted much attention to issues regarding the processes used to demarcate performance categories and the consequences of drawing those lines in high-stakes situations. Content standards have sometimes become embroiled in debates over “canons” of knowledge and the appropriateness and implications of designating certain facts, subjects, and skills as essential for a diverse population of students (Diegmueller, 1994). Both areas, then, would benefit from careful, empirically grounded research to substantiate the interpretations and uses of standards-based tests.

For my study, I focus on issues related to the content standards underlying the competency broadly understood as reading ability. Reading ability has long been identified and used as a key component of standardized educational tests in the United States. What “reading ability” actually denotes as a theoretical concept, and how it can best be measured with a standardized instrument, are questions that continue to find various answers in the theoretical literature and in the design of large-scale assessments.

To the extent that test results are used as the basis for decisions about the performance of schools, districts, states, and individual students, it is important that the construct of reading ability, as operationalized in those tests, be subject to careful empirical scrutiny, particularly with respect to the guidelines set forth in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Linn, 2002b). In particular, as states and districts integrate such tests into efforts to adopt standards-based accountability systems, the delineation of standards in content areas (such as reading) will likely have a major impact on curriculum design and instruction. Further, the use of standards to assess student performance and diagnose areas of strength and weakness means that a construct such as reading ability will be defined, for all intents and purposes, by the scores (and subscores) that each student earns.

The terms *reading* and *reading comprehension* are sometimes used to refer to the same ability, and sometimes to different ones, so clarification of my use of these terms is in order. Throughout this study, I follow a conventional distinction by using the terms *reading* and *reading ability* to refer to the entire process of decoding and constructing meaning from printed text. I use the term *reading comprehension* to refer primarily to the meaning construction aspect of reading; thus, reading comprehension refers to a part of the entire reading process (Lyon, 2002; Pearson, Barr, Kamil, & Mosenthal, 1984; Singer, Ruddell, & Ruddell, 1994; Smith, 1994). Based on this distinction, aspects of reading that would fall outside of reading comprehension include the visual processes involved in decoding text and the psychological process of converting text symbols to meaningful units of sounds, words, and sentences. This distinction does not represent a clear line between the two terms. For example, whether vocabulary is better understood as an aspect of reading or reading comprehension depends, to some extent, on one's theoretical perspective on how meaning is constructed during the reading process. Nevertheless, this distinction provides us with a useful way to more specifically discuss such theoretical perspectives. It should be noted, however, that where these terms are used specifically in the research covered in this study, I use the terms as originally used in that research.

The purposes of my study, which is based on the results of the 2002 Hawai'i State Assessment (HSA, formerly known as the Hawai'i Content and Performance Standards II State Assessment), are as follows:

1. To contribute to the existing body of knowledge on the validation of large-scale, standards-based tests, especially with respect to content standards.
2. To contribute to the existing body of knowledge on the construct of reading ability, especially as it is operationalized in standards-based tests.
3. To provide test validation information on the HSA to the State of Hawai'i Department of Education.

The Hawai'i State Assessment Reading Test

Background of the Hawai'i State Assessment

The Hawai'i State Assessment (HSA) is the state of Hawai'i's standards-based educational program. The program officially began with the creation of the Hawai'i Commission on Performance Standards by the State Legislature in 1991. In 1994, that commission published the Hawai'i Content and Performance Standards, which were widely known as the "Blue Book." In the same year, the State Legislature required the State Board of Education to appoint a Performance Standards Review Commission to review the Blue Book standards every four years, beginning in the 1997-98 school year. The commission's first report, issued in 1999, recommended significant changes to the Blue Book standards. Those recommendations resulted in revised standards, which were adopted by the State Board of Education in 1999 and were now known as the Hawai'i Content and Performance Standards, Second Edition, or HCPS II. The HCPS II underwent further study and revision, most significantly by a company under contract to the Hawai'i State Auditor, which had been requested by the State Legislature to examine HCPS II and make sure that its standards were in line with nationally comparable standards (Hawai'i State Auditor, 2001). The HCPS II standards guided the development of the State Assessment Program, the large-scale, standardized test designed to measure student achievement using the HCPS II standards. The State Assessment Program was first administered statewide in spring 2002. It is from that 2002 test administration that the data for this study is drawn. (Hawai'i Department of Education Office of Accountability and School Instructional Support/School Renewal Group, 1999; Hawai'i State Performance Standards Review Commission, 2003)

Framework of the HSA Reading Standards

The rationale for the HSA reading standards is presented in the Hawai'i State Department of Education's (DOE) *Curriculum Framework for Language Arts* (Hawai'i Department of Education Office of Curriculum Instruction and Student Support /

Instructional Services Branch, 2003), which also includes the standards for writing. The *Framework* provides "the theoretical and philosophical bases, grounded in sound research, upon which the content standards, benchmarks, and performance indicators were developed" (p. v). Thus, the *Framework* provides the broader context for understanding the HSA reading standards as the foundation for "curriculum, instruction, and assessment," which "are connected and must be aligned" (p. viii). The importance of the language arts standards is evident in the following statements contained in the *Framework*: "The Language Arts standards are derived from the goals of the Language Arts Program. They are the centerpiece of the Language Arts Program. They conceptualize how the Language Arts can be framed for assessment and instruction. They define what all students should know and be able to do with language" (p. 20).

Reading is one of five "areas of emphasis" within the language arts—including writing, oral communication, literature, and language study—that are "interconnected" to each other and to "personal knowledge, to schooling or technical knowledge, and to social or community knowledge" (p. 2). The role and value of reading are stated as follows: "*Reading provokes thought and reflection, allows readers to create and explore new ideas, and connects people to each other and to the world*" (p. 2, emphasis in original). Reading is defined as "a complex process of making sense of text and constructing meaning," and as a "recursive process" (p. 2).

The reading content standards, as well as those of the other subject areas of the language arts, were developed according to five guidelines adopted from the U.S. Department of Education and the Council of Chief State School Officers. These guidelines state that: (a) content standards should focus on the essential aspects of the subject and be based on solid scholarship, (b) content standards should be clear and applicable to educational practice, (c) the number of content standards should be "few," and each should be "bold and brief" in conveying the essential aspect of the standard, (d) content standards should be developed by consensus of stakeholders, and (e) content standards should represent goals, not current states, of learning (p. 12). Other sources for the criteria used to develop and revise the standards include the Council for Basic Education and national standards documents, which were not further specified (Hawai'i State Auditor, 2001).

The language arts content standards, which include the reading standards, are organized into four "strands." There is no stated definition of "strand," but based on the way the term is used it can be viewed as designating skills or capabilities that good readers possess. The DOE notes that although these strands are presented as separate and distinct for the sake of better understanding the standards, in actuality they are "intricately interwoven and constantly interacting" (Hawai'i Department of Education Office of Curriculum Instruction and Student Support / Instructional Services Branch, 2003, p. 12).

The strands and their descriptions are as follows:

1. *Range*—"Read a range of literary and informative texts for a variety of purposes including those students set for themselves."
2. *Processes*—"Develop and use strategies within the reading processes to construct meaning."

3. *Conventions and Skills*—"Develop and apply an understanding of the conventions of language and texts to construct meaning."
4. *Response and Rhetoric*—"Using individual reflection and group interaction, comprehend and respond to texts from a range of stances: personal, critical, and creative." (Hawai'i Department of Education Office of Curriculum Instruction and Student Support / Instructional Services Branch, 2003, pp. 14-16).

Three of the language arts strands—Processes, Conventions and Skills, and Response and Rhetoric—constitute the reading content standards, and are used as the basis for aligning the items in the HSA Reading test and for reporting a student's subscores on the test. The Range strand is incorporated into the HSA test in a different way. Rather than individual test items being aligned with the Range strand, the reading passages upon which test items are based are distributed among three different text types: literary, informational, and functional. This classification is explained in further detail below.

Structure of the HSA Reading Test

The complete HSA is composed of two tests: (a) The Stanford Achievement Test, Ninth Edition (SAT-9) Abbreviated. The SAT-9 Abbreviated consisted of a subset of items of the complete SAT-9, which were selected to enable comparability of scores on the SAT-9 Abbreviated and the complete SAT-9; and (b) The standards-based tests of reading, writing, and mathematics developed by the DOE. The complete HSA was administered to students in seven sections:

1. The Reading Comprehension Subtest of the SAT-9 Abbreviated test.

2. The Mathematics Problem Solving Subtest of the SAT-9 Abbreviated test.

3-4. Two standards-based reading segments of the HSA.

5. The writing assessment segment of the HSA.

6-7. Two standards-based mathematics segments of the HSA.

Thus, there were three segments (segments 1, 3, and 4) of the 2002 HSA that assessed reading ability (Hawaii State Department of Education, 2001).

Components of the HSA Reading Score

There are two main scoring components of the HSA Reading test. The first reflects the student's performance on all of the SAT-9 reading comprehension items. This score provides the basis for assessment of a student's reading performance against national SAT-9 norms. The second component reflects the student's performance on all of the standards-based, DOE-developed items, plus selected items from the SAT-9 segment that have been pre-identified by the DOE as being aligned with the HSA content standards for reading. Thus, a student's HSA reading score consists of items from both the SAT-9 Abbreviated and the DOE segments that shared a common trait in being aligned with the HSA content standards for reading.

The items that make up the complete HSA test can be grouped in several ways. Hereafter, I will use the following terms to denote the various groups described below:

1. *DOE items*: The questions developed by the DOE, all of which are aligned with the DOE reading content standards. On the 2002 test

administration, there were 37 and 36 DOE items on the Grade 8 and Grade 10 tests, respectively.

2. *SAT-9*: The SAT-9 Abbreviated Reading Comprehension Subtest, composed of 30 items. A student's score on the SAT-9 items is used to provide reading competency information relative to national norms. The SAT-9 items can be further divided into two groups:
    a. *Aligned SAT-9 items*: The subset of SAT-9 Abbreviated questions identified by the DOE as being aligned with its reading standards. These questions are counted in the student's HSA reading score. For Grade 8, there were 10 aligned items; for Grade 10, there were 16 aligned items.
    b. *Nonaligned SAT-9 items*: The remaining SAT-9 Abbreviated questions not aligned with the DOE reading standards, and not counted in the student's HSA reading score. For Grade 8, there were 20 nonaligned items; for Grade 10, there were 10 nonaligned items.
3. *HSA Reading test*: All of the reading test items given to students, composed of all of the SAT-9 Abbreviated items and all of the DOE items. Students were presented with a total of 67 items on the Grade 8 test, and 66 items on the Grade 10 test.
4. *HSA Reading score*: A student's HSA Reading score is the sum of (a) the aligned SAT-9 items, and (b) the DOE items. All of these items are

aligned with the DOE reading content standards. The Grade 8 reading score was composed of 47 items, and the Grade 10 score was composed of 52 items.

Table 1 summarizes the breakdown of the HSA Reading test items into alignment and content strands categories.

Table 1

Summary of HSA Reading Test Items

Source of Item	Aligned Items				Nonaligned Items	Total Items
	Comp. Processes	Conv. & Skills	Response	Total Aligned		
Grade 8						
SAT 9	1	2	7	10	20	30
DOE	15	11	11	37	0	37
Total	16	13	18	47	20	67
Grade 10						
SAT 9	1	4	11	16	14	30
DOE	12	11	13	36	0	36
Total	13	15	24	52	14	66

Proficiency Levels of the HSA Reading Test

The scores obtained by the HSA Reading test are used to generate three of the four sections of both the English Language Arts student report and summary report. (The fourth section is based on scores obtained by the writing test of the HSA.) The student report provides information on the performance of the individual student. This information is organized into three areas: (a) raw and percentile rank scores for the SAT-9 test; (b) raw and percent-of-total scores for each of the three reading strands with which the individual test items are aligned; and (c) identification of the performance level into which the student has been placed on the basis of the HSA score, and a narrative description of that standard. The summary report provides information on the performance of an individual class. This information is also organized into three areas that differ somewhat from those on the student report: (a) the average HSA reading (raw) scores for the class, school, district, and state, and the percentile rank of the class' average SAT-9 score relative to national SAT-9 norms; (b) the average score of the class on each of the three reading strands; and (c) the number and percent of students in the class who placed into each of the four proficiency levels (Hawai'i Department of Education, 2001, 2003)

A student's HSA Reading score items are used to place the student into one of four standards-based performance or proficiency levels: (a) well below proficiency, (b) approaches proficiency, (c) meets proficiency, and (d) exceeds proficiency (Hawai'i Department of Education, 2003). As discussed above, these performance levels are reported for both the individual student and for the student's class.

Classification of the HSA Reading Test Items

The basic format of the HSA Reading test as presented to students consisted of a reading passage followed by a series of questions. (A sample of this format is included in the Appendix.) Each of those questions can be classified along the following dimensions:

1. Alignment: Of the four language arts strands described in the *Curriculum Framework for Language Arts*, three are adapted specifically for classifying reading questions. These three reading strands are (a) Comprehension Processes, (b) Conventions and Skills, and (c) Response. (The fourth strand, Range, is incorporated in the passage type dimension, as explained on page 8.) Thus, each item is used as an indicator for, and is aligned with, one (and only one) of the three reading strands. As noted above, all DOE items are aligned with one of the standards, while only select SAT-9 items are so aligned.
2. Response Format: This refers to the type of response required by an item. Response formats include multiple-choice, open-ended, or extended response. (Note: A fourth format, short-answer, was not used in the 8th- and 10th-grade reading tests.) Multiple-choice (MC) items ask students to select a correct answer from four alternatives. Open-ended items ask students to respond to a prompt with an answer that could range from a few sentences to a paragraph. Extended response items ask students to respond to a prompt with an answer ranging from a few paragraphs to several pages.
3. Passage Type: This refers to the genre of the passage on which the item is based. The three passage types are literary, informational, and functional. Literary passages are short works or selections of fiction. Informational passages are designed to elicit the kind of reading that a student generally

> does for textbooks and other academic material for which the primary purpose is to gain information. Functional passages are designed to assess the student's ability to follow text instructions. The use of different passage types reflects the language arts standard of range.

Table 2 lists the complete HSA reading test items for the 8th and 10th grades, and indicates the categories within which each item can be classified.

Permission to Use Data

The DOE has granted me permission to use the data for the purpose of conducting this study. The permission form includes a confidentiality agreement, in which I agree to maintain the privacy of the students whose scores compose the data and to use the data only for the purpose granted by the DOE's permission. Hence, for this study all analyses were conducted on data aggregations that cut across identifiable groups. No analysis in this study constructed or used any variables that were based on districts, complexes, schools, classrooms, teachers, or individual students, and no attempt was made to identify the effects of, or attribute results to, those groups.

Table 2

HSA Reading Test Item Map

Grade 8				Grade 10			
Item	**Strand**	**Format**	**Type**	**Item**	**Strand**	**Format**	**Type**
SAT01	RE	MC	L	SAT01		MC	L
SAT02	CS	MC	L	SAT02	RE	MC	L
SAT03		MC	L	SAT03		MC	L
SAT04		MC	L	SAT04	RE	MC	L
SAT05		MC	L	SAT05	RE	MC	L
SAT06	CP	MC	I	SAT06		MC	L
SAT07		MC	I	SAT07		MC	L
SAT08	RE	MC	I	SAT08		MC	L
SAT09	RE	MC	I	SAT09	CS	MC	L
SAT10		MC	I	SAT10	RE	MC	I
SAT11		MC	F	SAT11		MC	I
SAT12		MC	F	SAT12	CS	MC	I
SAT13		MC	F	SAT13	CS	MC	I
SAT14		MC	F	SAT14	RE	MC	I
SAT15		MC	F	SAT15	RE	MC	I
SAT16		MC	I	SAT16		MC	F
SAT17		MC	I	SAT17		MC	F
SAT18		MC	I	SAT18		MC	F
SAT19		MC	I	SAT19	RE	MC	F
SAT20		MC	I	SAT20		MC	F
SAT21		MC	L	SAT21	RE	MC	F
SAT22	RE	MC	L	SAT22	RE	MC	F
SAT23		MC	L	SAT23		MC	F
SAT24		MC	L	SAT24		MC	I
SAT25	RE	MC	L	SAT25		MC	I
SAT26		MC	F	SAT26	RE	MC	I
SAT27		MC	F	SAT27		MC	I
SAT28	RE	MC	F	SAT28	RE	MC	I
SAT29	CS	MC	F	SAT29	CS	MC	I
SAT30	RE	MC	F	SAT30	CP	MC	I
HSA01	CP	MC	L	HSA01	CP	MC	L
HSA02	CS	MC	L	HSA02	CS	MC	L
HSA03	RE	MC	L	HSA03	CS	MC	L
HSA04	CP	MC	L	HSA04	CS	MC	L
HSA05	CP	MC	L	HSA05	RE	OE	L
HSA06	CP	OE	L	HSA06	CS	MC	I
HSA07	RE	EX	L	HSA07	CS	MC	I
HSA08	CP	MC	F	HSA08	CS	MC	I
HSA09	CS	MC	F	HSA09	CS	MC	I
HSA10	RE	MC	F	HSA10	CP	OE	I
HSA11	CS	MC	F	HSA11	RE	OE	I
HSA12	CS	MC	F	HSA12	CP	MC	L
HSA13	RE	OE	F	HSA13	CP	MC	L
HSA14	RE	OE	F	HSA14	CS	MC	L
HSA15	CS	MC	L	HSA15	RE	MC	L

Table 2 (continued)

HSA Reading Test Item Map

Grade 8				Grade 10			
Item	**Strand**	**Format**	**Type**	**Item**	**Strand**	**Format**	**Type**
HSA16	CS	MC	L	HSA16	RE	MC	L
HSA17	CS	MC	L	HSA17	CP	MC	I
HSA18	RE	MC	L	HSA18	CP	MC	I
HSA19	RE	OE	L	HSA19	CP	MC	I
HSA20	CP	MC	L	HSA20	CS	MC	I
HSA21	CP	MC	L	HSA21	RE	MC	I
HSA22	RE	MC	L	HSA22	RE	MC	I
HSA23	CP	OE	L	HSA23	RE	EX	I
HSA24	CP	MC	F	HSA24	RE	OE	I
HSA25	CP	MC	F	HSA25	CP	MC	I
HSA26	CP	MC	F	HSA26	CP	MC	I
HSA27	CP	MC	F	HSA27	CS	MC	I
HSA28	CP	MC	F	HSA28	CS	MC	I
HSA29	CS	MC	F	HSA29	RE	MC	I
HSA30	CS	MC	F	HSA30	RE	EX	I
HSA31	RE	EX	F	HSA31	CP	MC	F
HSA32	RE	MC	I	HSA32	CP	MC	F
HSA33	CP	MC	I	HSA33	RE	MC	F
HSA34	CS	MC	I	HSA34	RE	MC	F
HSA35	CS	MC	I	HSA35	CP	OE	F
HSA36	CP	MC	I	HSA36	RE	EX	F
HSA37	RE	OE	I				

Note. Abbreviations used in table:

Content Strand: CP = comprehension processes, CS = conventions and skills, RE = response, blank = nonaligned

Format: MC = multiple-choice, OE = open-ended, EX = extended response

Type: L = literary, I = informational, F = functional

CHAPTER 2. REVIEW OF THE LITERATURE

Construct Validation of Large-Scale, Standards-Based Tests

The most recent edition of the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999) begins with the assertion that validity "refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests" (p. 9). Since one of the proposed uses of the HSA is to determine the proficiency level of students for a construct identified as "reading" as well as indicating their performance on each of the three components or "strands" of reading, it is useful to consider some of the theoretical issues involved in construct validation. Geisinger (1992) succinctly charts the evolution of the concept of validation from the 1954 publication of the American Psychological Association's *Technical Recommendations for Psychological Tests and Diagnostic Techniques*, through the 1985 *Standards for Educational and Psychological Tests and Manuals*. Geisinger points out that there has been clear movement towards a unified concept of construct validity that subsumes traditional ideas of content and criterion validity. While this has led some researchers to downplay the importance of content- and criterion-related validation, he argues that the two provide key support for broader construct validation efforts, especially with respect to tests in which "the domain delineations and definitions of success [are] determined by expert judgments. Thus, although content-related validation does not consider the responses of test takers in its evaluation of a test and therefore has been removed by some from the set of validation strategies, criterion-related validation determines validity by assessing the degree of correlation with presumably content-valid criteria" (p. 208). This point is particularly relevant to an analysis of a test like the HSA where the content standards and the test items are developed through a process of expert opinion and review.

Messick's (1989) discussion of validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (p. 13, emphases in original) provides an extensive argument for basing validity assessments on the uses of test scores. Linn's (2002a) elaboration of the implications of this concept of validity for standards-based assessment covers several topics that are of much relevance to this study, including changing conceptions of content areas and the manner in which performance standards are implemented in tests of those areas.

Validity issues relating to the use of test scores for normative and curriculum-specific assessment explored in Linn and Hambleton (1991) are relevant to the HSA, which is designed to provide both normative and curriculum-specific information to parents and educators. The many potential threats to validly drawing both normative and standards-based conclusions from the same data at the national level are discussed by Linn (1998). Linn and Hambleton (1991) note that making valid inferences from large-scale tests is complicated by the many purposes for which such tests are designed; often, "a testing program designed to serve well one of these purposes may do a relatively poor job of satisfying another expectation" (p. 186). The authors argue that caution is required when constructing tests to provide meaningful information for both normative and standards-based purposes. Ideally, students would be given multiple tests, each designed for a specific purpose. Doing so, however, would be financially burdensome, and likely allocate an unreasonable amount of instructional time to testing; thus, "there are strong pressures for the development of efficient testing systems that can serve multiple purposes simultaneously" (p. 186). The practical demands for testing efficiency, however, raise a fundamental question: "Can a test serve multiple purposes and retain an adequate level of validity for each purpose?" (p. 194).

An important aspect of validity investigations within a standards-based paradigm is alignment: the extent to which the test items themselves accurately represent the standards that test-takers are being measured against. "For accurate inferences to be made about student achievement and growth over time, these [standards-based] assessments must measure the knowledge and skills deemed valuable and described in policy documents such as state content standards. From this perspective, alignment has both content and consequential validity implications" (Bhola, Impara, & Buckendahl, 2003, p. 21). Bhola et al. (2003) review four models of assessing the degree and quality of alignment between test and standards, and adapt Webb's (1997) model in their study of the reading/writing and mathematics content standards of Nebraska's statewide test. The alignment models discussed in Bhola et al. (2003) provide a useful context for thinking about how test items can be analyzed in terms of how they represent both the breadth of the content area and the depth of thinking by the test taker. In a comprehensive review of elementary school level reading standards of all fifty states, Wixon, Fisk, Dutro, and McDaniel (2002) found that alignment was interpreted quite differently by different states. Although the processes used by the states in aligning standards with tests were fairly clear-cut, assessing the extent of alignment was complicated by a variety of differing interpretations of adequate coverage of standards.

Components of Reading

Terms and Approaches Used in Studying the Problem

In a review of reading models dating from the mid-1960s to the mid-1980s, Samuels and Kamil (1984) noted that one of the difficulties of comparing such models lay in the different knowledge bases and theoretical assumptions that each is built on, reflecting the influence of scientific philosophies extant at the time. Samuels and Kamil focus on models based in information processing (IP) theory, but make clear that even within this framework there are large contrasts among the models. Similarly, current understandings of reading components come from a variety of perspectives both within and across theoretical frameworks. For this reason, the terms used to conceptualize and define reading components, including "component," "dimension," "subskill," and "factor," reflect various theoretical frameworks, perspectives, methodological choices, and what Brennan (1998) has identified as a fair amount of confusion over the constructs the terms are intended to denote. In general, though, the terms "component" and "subskill" are used to refer to discrete processes and capabilities within an overall reading process conceptualized within IP and developmental frameworks; the term "dimension" is often used within an item response theory (IRT) framework that emphasizes the characteristics of test questions; and the term "factor" is generally used to refer to the constructs that emerge from latent factor analyses. Thus, each of the terms indicates, to some degree, the relative emphasis placed on the hypothesized model of reading components versus the method used to discern those components. In the discussion that follows, the terms are used as they were originally used in the studies being covered, and the term "component" is used when making reference to the general concept of subdivisions within reading competency. Similarly, the terms "decoding," "word recognition," "phonological," and "word building" are often used to refer to the processes and skills involved in perceiving, recognizing, and representing text for meaningful construction by the reader. The term "comprehension" is used to refer to the processes and skills involved in constructing meaning, making inferences, and monitoring understanding. These two general classes of skills are often referred to as "lower level" and "higher level," respectively, although as will be discussed below these designations, at least in terms of their common-use implications, may not be entirely appropriate. These terms will also be used as originally used in the studies discussed.

The discussion of the literature that follows is organized around three perspectives on models of reading within a general cognitive theoretical framework—developmental, information processing, and holistic—as well as actual applications of reading models in large-scale testing. The focus in this discussion is on several key issues that bear on the problem of measuring components with a large-scale test. Subsequently, I consider applied approaches to reading components that inform current large-scale assessments.

Developmental or Language Ability Level Perspectives

One approach to reading components is to treat them within the context of the learner's cognitive or language ability development. From this perspective, components and the relationships among them are generally treated as dynamic and evolving rather than static. Key assessment issues involve the use of developmentally appropriate methods to measure reading ability, and the relationship between reading ability and other cognitive processes.

One implication of a developmental perspective is that specific components may be theoretically and diagnostically useful only during certain periods in a reader's development. Stanovich's (2000) review of critical themes in cognitive developmental approaches to reading indicates that research evidence supports a model of reading as composed of two general processes that change in importance over time and across individuals of different reading skill levels: a word recognition process and a comprehension process. He notes that theories of cognitive resource constraints, automaticity, and modularity, have successively shaped our current understanding of how individuals develop as readers with respect to their cognitive abilities. In particular, Stanovich argues for a modularized view of the word recognition process in good readers as independent of higher level, context-related knowledge and skills. This modularity, he notes, refers to the informational self-sufficiency of the word recognition process, and not, as had been earlier hypothesized, to its being free of cognitive resources or attention. Thus, poor readers, or those at an earlier stage of development, rely on contextual cues and other aspects of the reading environment to assist them with word recognition, thereby reducing the resources available to them for comprehension.

This perspective informs Stanovich's (1986/2004) well-known discussion of the "Matthew effects" of reading difficulties, in which he argues that in identifying specific reading problems one must account for the reciprocal relationship between cognitive development and reading skills. Early reading problems lead to problems in related cognitive skills, which in turn prevent reading skill development, and left unchecked this process becomes a downward spiral. Conversely, in normal reading development the nature of reading changes as the reader develops, with skills becoming more interdependent and, hence, more difficult to isolate and specify as sources of reading difficulty. Thus, he notes that the empirical evidence for specific reading skills beyond a certain level of development is weak, but urges that such a line of research is worth pursuing. This view is supported by Alderson (2000), who suggests that "component skills approaches may be valid and justified for beginning, weak, dyslexic or low-level second-language readers, but not for more advanced readers" (p. 97). However, he notes that the evidence for this position is far from conclusive, with some studies finding very high correlations among components (suggesting one general reading skill) even among these populations.

The changing relationship between word recognition skills and comprehension skills is a key concern of researchers working within a developmental framework. Rupley, Willson, and Nichols (1998) find support for a modified version of Carver's (1993) theory of general reading ability, or "rauding" ability. Rauding components posited by Carver include a general intelligence factor called cognitive power, a cognitive speed component that reflects naming and identification speed, listening comprehension, and word pronunciation ability. Rauding theory, in turn, is an extension of Gough's simple view theory (Gough, Hoover, & Peterson, 1996) that takes into account the changes in reading development in the early years of schooling. In the simple view model, reading consists of a decoding component and a general comprehension component. Rupley et al. argue that the simple view, as originally conceptualized, is weak in addressing other factors affecting reading development in elementary school children. Following Carver, they find evidence indicating that for students in grades 1 through 6 word recognition plays a key role in reading ability, but that role diminishes as students grow older. Thus, Rupley et al. argue that decoding skill is important largely during the younger years, and as a reader ages the relationship between decoding and comprehension decreases.

Research into second-language (L2) learning also deals with developmental issues, but in relation to the learner's level of proficiency in the target language rather than his or her psychosocial or cognitive development. Carlo and Sylvester's (1996) synthesis of the literature on L2 reading research is organized around Perfetti's (1988) verbal efficiency model of reading. According to the verbal efficiency model, reading processes can be divided into local text and text modeling processes, which refer, respectively, to lower level processes used to retrieve information from the specific text being read (such as syntax, letter and word recognition, and word access) and higher level processes that use the output of the local text processes to generate meaning. In this view, local text processes must be operating sufficiently well in order for reading to occur, since they create the units out of which meaning is built; thus, developmentally speaking, local text processes precede text modeling processes. From the perspective of L2 reading assessment, then, it is important to isolate local text processes in order to get an accurate picture of the student's ability level.

With respect to the assessment of reading ability and its hypothesized components, then, these studies suggest that any such components must be conceptualized within a framework that accounts for changes in the reading process as readers develop. However, developmental changes in cognition and reading may have varying effects on efforts to assess components. For example, in a diverse group of readers evidence of comprehension subskills demonstrated by proficient readers may be difficult to disentangle from results reflecting difficulty with word recognition processes in less developed readers. On the other hand, evidence for components may be more evident in such a group, if those components are designed to reflect skills that are distinguishable in less developed readers because of their need to devote higher level thinking and contextual cues to word recognition processes.

Information Processing Perspectives

In a sense, the very concept of reading as consisting of components is very much within an IP paradigm (Grabe & Stoller, 2002). And within that paradigm, reading—like other kinds of cognitive activities—is, by definition, an activity involving the various resources, components, and processes of the mind, such as long-term memory and working memory. Research that adopts an information processing perspective treats components as discrete subprocesses or capabilities within a larger process of comprehension. LaBerge and Samuels, in their influential article (1974) on the automaticity of reading, state that "during the execution of a complex skill, it is necessary to coordinate many component processes within a very short period of time. . . . Therefore, one of the prime issues in the study of a complex skill such as reading is to determine how the processing of component subskills becomes automatic" (LaBerge & Samuels, 1974, quoted in Stanovich, 2000, p. 222). Similarly, Fuchs, Fuchs, Hosp and Jenkins (2001) describe reading as "a complex performance that requires simultaneous coordination across many tasks. To achieve simultaneous coordination across tasks, instantaneous execution of component skills is required" (p. 239). Thus, IP-informed research is generally concerned with identifying components and specifying their relationship to each other and to general reading ability (Carr, Brown, Vavrus, & Evans, 1990).

As Hannon and Daneman (2001) note in contrasting single- and multiple-component theories of reading, however, many IP-based theories have argued for a single component model of reading ability. There is little agreement on what that single component is; answers have included word recognition skill, one of several hypothesized higher level skills of text and knowledge integration, and working memory capacity. Such single-component models have been criticized as inadequate because they do not account for the wide range of cognitive and linguistic abilities that have been found to be correlated with reading, and thus do not facilitate progress in understanding of individual differences in reading. Hannon and Daneman (2001) argue that multicomponent models, particularly those that focus on higher level processes, more effectively account for variance on standardized reading tests that have been shown to be associated with various cognitive skills.

There are differing opinions on the importance and centrality of decoding skill, which is often identified as being the most empirically justified of any component in a single-component model, and is also prominent in many multiple-component models. In summarizing a collection of IP research on reading components, Levy and Carr (1990), while acknowledging that phonological skills appeared to be central to reading ability for both children and adults, concluded that a "single-factor phonological theory" was clearly unacceptable (p. 434). Levy and Carr point out that it is not only the contribution of higher level skills that is persuasive on this point, but the "patterns of intercorrelation" among those skills (p. 434). A contrasting perspective is offered by Chard, Simmons, and Kameenui (1995) in their overview of reading research. Chard et al. specify four "points of convergence" at which the evidence indicates that word recognition plays the pivotal role in reading ability. On the basis of this evidence, they liken word recognition to the fibers of a metaphorical rope representing reading ability.

Gough, Hoover, and Peterson's (1996) "simple view" is, as its name suggests, simple in enumerating just two component skills. It is, however, explicitly a component model of reading, offered against the perspective of reading as a general, undifferentiated skill. In the simple view model, only one of the two components, decoding, is unique to reading, while the other, general comprehension, applies to both printed text and verbal input. Gough et al. argue that researchers' inability to empirically isolate components results from using methods that do not take this basic division into account, and rely on measures that are all based on decoding. By separating the two skills, Gough et al. found evidence that reading ability is a multiplicative product of decoding skill and comprehension skill, such that low ability in either will result in low reading ability. Catts, Hogan, and Fey (2003) found that a reader classification strategy (the Reading Component Model) based on the simple view was useful in grouping and profiling less able readers in ways that facilitated appropriate intervention, in contrast to strategies that utilized global measures of cognitive ability. Although the Catts et al. study is based on a sample of younger students (kindergarten through fourth grades), it provides a measure of empirical support for viewing comprehension and decoding skills as independent, and for emphasizing their relationship not with cognitive development but with reading skill.

Beyond the two components of the simple view model, the number and types of components are wide-ranging. Cunningham, Stanovich, and Wilson (1990) used a variety of tests to measure 21 variables in the areas of comprehension (including listening comprehension), working memory, phonological skills, and reading habits for 90 college students. Using factor analysis, they arrived at a three-factor model to account for differences between good and poor readers: reading comprehension, word recognition, and general verbal comprehension. It is notable that these factors are very similar to the components of the simple view model, especially with regard to the continued importance of decoding skill in adults.

Correlations among twelve variables similar to those used by Cunningham et al. are the basis for Cain, Bryant, and Oakhill's (2004) study of 102 elementary school-age children. Cain et al. found that working memory and three higher level comprehension skills—the ability to make inferences from the text, the ability to self-monitor comprehension, and knowledge of narrative text structure—accounted for unique variance over and above decoding skills and verbal ability. Hannon and Daneman (2001), using a modified version of an instrument developed by Potts and Peterson (1985), identified four reading comprehension components: the ability to retrieve prior knowledge from long-term memory, the ability to integrate that knowledge into newly acquired information from the text, the ability to make inferences on the basis of that integration, and the ability to recall newly acquired text information.

The relationship among reading components is an issue that has important implications for reading assessment. Both Gough et al.'s simple view of reading and the modified version supported by Rupley et al. provide for a degree of independence between decoding and comprehension. From this perspective, both skills develop in parallel, and although the modified model indicates a more important role for decoding in early school years, neither skill is seen as being more basic than the other. There is no hierarchical relationship between the two; decoding is not a "lower level" skill in the sense that it is required for comprehension, and vice versa.

A contrasting perspective is provided by Cain et al. They find evidence to support a "bottom-up" view of reading, according to which lower level skills, such as decoding, constitute an essential base for higher level skills required for comprehension. Although there is evidence that some higher level component skills contributed unique variance to reading ability, in general Cain et al.'s results confirm the theoretical position that reading is a "bottom-up" process, in which problems with lower level skills will impair higher level skill functioning. Similarly, Perfetti, Marron, and Foltz (1996) examine areas of reading failure in three subjects and conclude that the burden of proof remains on those who claim that higher level comprehension problems can exist in the absence of lower level problems. McCandliss, Beck, Sandak, and Perfetti (2003), while acknowledging that improvements in decoding skill do not always result in benefits to comprehension skill, find support for decoding as the primary reading component.

Context-based and Holistic Perspectives

Frameworks that emphasize the contexts and holistic aspects of reading view it as an integrated and fluid process in which components, even if they do exist, have very little practical usefulness in understanding individual differences in reading or in diagnosing reading difficulties. (This perspective is sometimes labeled "schema-based," but in this discussion I use the terms "holistic" and "context-based" to emphasize those aspects of this perspective and to distinguish it from early models (e.g., Anderson & Pearson, 1984) characterized by a modularized and diagrammatic approach. More recent explications of schema-based reading theory, such as Anderson (2004), adopt a perspective that is closer to the one discussed here.) Proponents of this view come from different perspectives on the nature of reading ability, but they agree that it should be assessed as a single skill. From one perspective, reading is understood as a complex and holistic process involving reader, text, and reading environment. Thus, the decomposition of reading into components serves to decontextualize reading in an artificial manner. Johnston (1984) argues that reading "does not consist of a set of discrete subskills . . . but involves the integration of a variety of declarative, procedural, and strategic knowledge in different ways, depending on the state of the comprehending system and the information available to it from various sources" (p. 160). From this perspective, reading research would be better served by focusing on reader-text interaction, the reader's prior knowledge, and the circumstances of the act of reading, rather than searching for evidence of discrete components. Properly designed test items would thus be constructed with two main dimensions in mind: the test taker's prior knowledge, and the central elements of the text that the item is assessing. "If one could know the extent and nature of relevant knowledge held by readers prior to their reading a passage, one would know much more about the nature of the task posed by questions following the text and the nature of the strategies that could be employed" (Johnston, 1984, p. 155). From this holistic view of reading, then, components are artifacts of theory inappropriately applied to the process of reading.

Other researchers question efforts to assess reading components because, while acknowledging that such components do or may exist, they argue that those components are simply too tightly integrated or difficult to measure in order for meaningful assessment to make use of them. Schwartz (1984), for example, notes that reading subskills tend to be "fairly broad" and "highly intercorrelated," and hence argues that "the possibility that there is only one general skill involved in comprehension or reasoning cannot at present be rejected" (p. 87). Alderson (1990) conducted a study in which expert judges were asked to identify and agree upon the reading skills that various test questions were assessing, and found that there was little agreement among the judges.

On the basis of this evidence, he argued that "even if there are separate skills in the reading process which one could identify by a rational process of analysis of one's own reading behaviour, it appears to be extremely difficult if not impossible to isolate them for the sake of testing or research" (p. 436, cited in Alderson, 2000, p. 49).

With respect to the importance of context in understanding reading ability, Stanovich (1994) has pointed out that reading theory "is quite interestingly bifurcated" (p. 264). On the one hand, research evidence strongly supports the important role of background knowledge and contextual factors in reading comprehension. In this context, constructivist interpretations of and instructional methods for reading improvement make good sense. On the other hand, there is also much evidence that word recognition processes are more appropriately developed by direct instructional methods. Reading is thus a "special type of constrained reasoning" (p. 264) that poses a challenge for instruction and assessment by making them account for two, seemingly divergent needs and approaches within a process that cannot easily be decomposed.

Constraints on Using Components to Inform Large-Scale Testing

To what extent are findings of reading components facilitated or constrained by the methods and frameworks discussed above? This is an important question, because it addresses the key distinction between the existence of reading components and our ability to measure their existence in a large-scale testing context. Approaches from a developmental or reading ability framework, as well as other frameworks that emphasize the hierarchical relationship between lower- and higher level components, suggest that proper assessment depends on the isolation of local text processes from text modeling processes, to use the distinction made by Perfetti (1988). Without accurate knowledge of a student's basic skills, we would have little information on which to base assessments of meaning construction and metacognitive skills.

The other issue that is raised by developmental frameworks is identification of the ages, or stages of reading ability, at which assessment of local text and text modeling processes are appropriate. If local text processes become progressively more difficult to discern as a reader develops, then at some point assessment of those processes may become unreliable. Similarly, if text modeling processes can only be meaningfully assessed relative to the development level of local text processes, then it is important to know the point at which assessment of those higher level skills can begin. Inadequate performance on items testing lower level skills might suggest that scores on higher level skills be discarded as uninterpretable (since the test taker has, by definition, displayed an inability to perform higher level skills). Or, test developers would need to design higher level items in such a way that certain types of errors would indicate deficiencies in lower level skills.

Although Gough et al.'s (1996) simple model is similar in outline to the developmental framework used by Carlo and Sylvester (1996), the independence of the decoding and comprehension components in the simple model suggest that it would be a more difficult framework for developing a large-scale assessment. Specifically, if comprehension must be assessed independently of decoding in order to be validly measured, then using a paper-and-pencil test must be limited to assessment of decoding skill. Gough et al.'s method for distinguishing between the effects of decoding and comprehension relied on a measure of verbal comprehension (“listening”) assessed by reading a story to a child and then asking questions about it. It is questionable whether this method of assessment can be accommodated in a large-scale test, even one using a multimedia format capable of delivering both auditory and text passages and questions, because the student’s response to the auditory prompts would likewise be spoken. In Rupley et al.’s (1998) modification of the simple view, this limitation would be less salient as the reader develops and decoding skills become less important in distinguishing reading ability. However, if we follow Gough et al. in positing a more complete separation of the two skills, then decoding can be a source of reading difficulty at any age or development level. Hence, proper assessment would require a paper-based test of decoding and a verbal test of comprehension at all age levels.

It should be emphasized that evidence for components in many of the studies discussed here relies on multiple measures. In the Cain et al. study (2004), for example, standardized instruments included the Neale Analysis of Reading Ability, Gates-MacGinitie Vocabulary subtest, and subtests from the Wechsler Intelligence Scale for Children; face-to-face assessments were used to measure working memory, inference making skills, comprehension monitoring, and knowledge of story structure. The four components in Hannon and Daneman’s (2001) model were measured with a computer-delivered instrument, and validated against the Nelson-Denny standardized comprehension test, the Mill Hill test of vocabulary knowledge, verbal analogies test items, a deductive reasoning test, and items from the analytic section of the Graduate Record Examination. Rupley et al. (1998) analyzed scores on the Kaufman Assessment

Battery for Children and the Kaufman Test of Educational Achievement. Cunningham et al. (1990) applied a factor analysis to data comprising 22 variables generated by ten different tasks. From a psychometric perspective, then, we might expect to find more evidence of different components in this type of analysis simply due to the effects of using multiple and diverse measuring instruments.

From a holistic perspective, assessment efforts are better directed towards the context of reading and the interaction between reader and text rather than towards isolation of components. Hence, the link between assessment method and components is relevant only to the extent that such a method would account for contextual variables. With respect to a test like the HSA, then, the key issue would not be whether the patterns of responses could provide evidence for components; rather, it would be whether those response patterns could be examined meaningfully within reading contexts provided by such variables as the circumstances of the assessment and the relationship between the subject matter of the test items and reader's prior knowledge. One example of a context factor derived from social learning theory is the self-efficacy of the reader, which has been shown to have significant effects on reading performance and on related factors such as motivation and perseverance, particularly in high-stakes testing situations (Guthrie, Wigfield, Metsala, & Cox, 2004; McCabe, 2003). Thus, there is no inherent reason that a test like the HSA could not provide useful assessment information, but its results would need to be fully contextualized in order to be validly used.

An interesting perspective on the relationship between component skills and reading ability is adopted by research in the area of emergent or early literacy. On one hand, this research stresses the holistic aspects of literacy development, focusing particularly on environmental and interpersonal factors that facilitate literacy development at home and in other circumstances prior to formal instruction. On the other hand, there is a strong concern for the developmental trajectory of literacy skills that need to be encouraged and facilitated at an early age, such as phonological and print awareness and oral language skills (Pullen & Justice, 2003). Research linking these components of early literacy to reading achievement in later years provides a framework in which component skills can be isolated for the sake of developing instructional strategies, but from a holistic approach to the reading context (Whitehurst & Lonigan, 1998).

Applied Frameworks

Another perspective is informed by the teaching of reading, and approaches reading components as skills necessary for reading a variety of texts in diverse situations. Research based on this perspective tends to be more concerned with macro-level processes and problems than with isolating specific processes and capabilities. Reading components are largely derived on the basis of expert judgment and experience, and their definitions are more clearly tied to instructional strategies that can be used to address deficiencies in those components.

Various stakeholder groups and instructional experts have developed models of reading that are intended to inform or assess content standards. A nationally applied model is used in the National Assessment of Educational Progress (NAEP) reading assessment, which the HSA reading standards closely resemble. The NAEP program administers tests to a national sample of 4th- and 8th-grade students, the data from which generates the "Nation's Report Card" on a number of subjects, including reading. It is primarily a norm-referenced test, designed to provide comparative data for states and regions. The rationale for the NAEP reading test is included in the NAEP Reading Framework, which "reflects the ideas of many diverse individuals and organizations involved in reading education. In developing the framework for the national assessment of reading, researchers, policymakers, teachers, business representatives, and other experts have specified behaviors of proficient readers who are active, strategic, knowledgeable, and motivated to read" (National Assessment Governing Board, 2003).

The NAEP reading test questions are classified into four "aspects" of reading: (a) forming a general understanding, (b) developing interpretations, (c) making reader/text connections, and (d) examining content and structure. These aspects are assessed across three "contexts" of reading: (a) reading for literary experience, (b) reading for information, and (c) reading to perform a task. Despite the similarity between the NAEP and HSA models, it is interesting to note that the NAEP aspects of reading, unlike the HSA strands, are explicitly non-diagnostic, designed only to measure "overall achievement." According to the National Assessment Governing Board (2002), "NAEP examines whether students can use multiple skills, not specific skills, to comprehend what they read." Nevertheless, the aspects represent more than just facets of a unitary reading skill, as they are intended to indicate discrete abilities, where "successfully mastering one aspect may not depend on successfully mastering any other aspect" (National Assessment Governing Board, 2003).

Other models of reading have been used as points of comparison for assessments of the HSA. In 2001, the Hawai'i State Auditor conducted an evaluation of the HSA (then known as the HCPS II) (Hawai'i State Auditor, 2001). The Auditor's evaluation of the HSA content standards was based on work done by a consultant, Mid-continent Research for Education and Learning (McREL). In turn, the McREL assessment was based on assessments of state standards by the American Federation of Teachers, the Fordham Foundation, and the Council for Basic Education (Hawai'i State Auditor, 2001, pp. 39-40). Based on those assessments, McREL identified five states as exemplars in setting content standards in the area of language arts: Arizona, California, Massachusetts, Virginia, and Wisconsin.

On the basis of its review of those states' standards, McREL developed a list of standards and benchmarks intended to "provide schools, districts, and states with a means for identifying the knowledge and skills that are most important for students to learn for the subject areas of language arts, mathematics, and science" (Kendall, Snyder, Schintgen, Wahlquist, & Marzano, 1999, p. 1). The three general reading standards specified by McREL ask readers to demonstrate "competence in the general skills and strategies" of the reading process, of reading a variety of literary texts, and of reading a variety of informational texts (Kendall et al., 1999). Each of these general standards includes several subskills. McREL's list clearly indicates that reading context is of primary importance, as two of the three standards are specific to text types. This suggests a two-dimensional perspective on reading consisting of competence in basic reading skills in a generic sense, and of skills that are specific to certain text situations.

One of the three reports used by McREL to develop its standards is authored by the Council for Basic Education (CBE). The CBE model of reading is presented in a document entitled *The Keys to Literacy* (Patton & Holmes, 2002). That document identifies phoneme awareness (the ability to connect sounds to print), fluency (reading with sufficient speed and accuracy that meaning construction is possible), and meaning construction as central skills in reading (Lyon, 2002). Meaning construction, or reading comprehension, depends most importantly on an adequate vocabulary, activation of background knowledge, an understanding of word relationships, writing conventions, and metacognitive skills to check and question the meanings obtained from the text (Beck & McKeown, 2002; Lyon, 2002). The CBE model of reading places a heavy emphasis on phoneme awareness as the basis for reading, and recommends extensive and early skill building for students.

The report issued by the Fordham Foundation (Stotsky, 1997) includes several criteria for evaluating state standards that provide the outlines of a model of reading. It states that standards should be organized to reflect the distinction between "higher-order knowledge and skillls from lower-order skills." It also states that reading standards should reflect the importance of reading for information across a range of different contexts; however, particular emphasis is placed on the full range of skills needed to appreciate literature. In general, the Fordham Foundation criteria reflect a focus on reading comprehension, with little attention paid to "lower-order skills" of the reading process. Like the CBE and McREL criteria, it notes the importance of reading across a range of contexts.

The third report that informs McREL's assessment of the HSA was developed by the American Federation of Teachers (American Federation of Teachers, 2001). The AFT report does not enumerate specific standards for content areas or performance. Instead, it evaluates state standards on criteria relating to the explicitness and detail of the standards, the specificity of standards to grades or grade ranges, the range of content areas covered by standards, and the degree to which assessments are based on the specific state standards.

The models of reading that underlie the criteria used by these organizations to evaluate state reading standards possess a relatively low degree of specificity, and if the McREL study is any indication of the information base upon which state-level, large-scale reading assessments are designed, it is no wonder that there is much diversity among them. That reading is composed of several components or skills is implicit in most of these models, but important questions regarding those components remain unaddressed: (a) Are components of reading discrete and, thus, separately testable? (b) How might these components be related with each other? Do they exist in a hierarchical relationship, with proficiency in "higher order" skills indicating (and dependent on) mastery of "lower order" skills? Or do they exist in a more equal relationship, with facility in one independent of that in another? (c) To what extent does the measurement of these components depend on text characteristics, such as the format, subject matter, and authenticity of the test question? These kinds of questions, which are critical to any effort to develop a standards-based reading test, remain for individual states and test developers to wrestle with.

Although the answers to these questions are implicit in the design of many large-scale tests, there is a lack of empirical research to substantiate the theoretical models behind the different approaches to reading embodied in those tests. Even a cursory review of state-level educational testing programs indicates that there is a lack of consensus on how to best understand and measure student reading ability (Florida Department of Education, 2001; Massachusetts Department of Education, 2001; Texas Education Agency, 2003; Washington Office of Superintendent of Public Instruction, 2000). Perhaps this diversity of reading models and assessment practices reflects basic problems with existing taxonomies of reading that Alderson (2000) has identified: (a) they are often based more on expert opinion and induction, rather than empirical evidence; (b) reading components tend to lack clear definition, and are often conceptually non-discrete; (c) it has proven difficult to link test items with individual components; and (4) analyses of test performance often fail to indicate the presence of components.

Comparison of State Test Structures and Content Standards

Table 3 summarizes test structure and content standards for the five standards-based state assessments identified by McREL (Kendall et al., 1999) as model reading assessment programs: the Arizona Academic Content Standards (Arizona Department of Education School Effectiveness Division, 2003), California Standardized Testing and Reporting Program (California Department of Education, 1998), Massachusetts Comprehensive Assessment System (Massachusetts Department of Education, 2001), Virginia English Standards of Learning Curriculum Framework (Virginia Department of Education, 2003), and the Wisconsin Model Academic Standards (Wisconsin Department of Public Instruction, 1998). Content standards are categorized into two main groups: those that pertain to reading components or subskills, and those that pertain to reading contexts or passage types. In general, the state assessments included appear to devote more attention to reading contexts and passage types, and less attention to isolating reading components. Content standards relating to reading contexts are more numerous and include more detail than those relating to components. Standards relating to reading components vary from those concerned with decoding skills (California) to metacognitive strategies (Wisconsin). Standards relating to reading contexts are more consistent across the states, with a relatively heavier emphasis on literary texts than other types.

Psychometric Evidence of Components from Large-Scale Tests

Studies of reading components that have looked primarily at student performance on large-scale tests have been less successful at isolating those components. In their analysis of the TOEFL reading comprehension test items, Schedl, Gordon, Carey, and Tang (1996) found little evidence for distinguishing between items designed to tap "higher level" reading skills, and general reading ability items that assessed vocabulary, syntax, and explicit information. However, the authors did find weak evidence that the tests were not unidimensional and that there was a "minor secondary factor" related to reading passage content or position. In an analysis of second language reading comprehension test results, Buck, Tatsuoka, and Kostin (1997) identified 24 "attributes" of reading ability that explained 97 percent of the total test score variance. However, their list of attributes includes many that are statements of general cognitive ability and

Table 3

Summary of Grade 8 Content Standards of Selected State Reading Tests

	Arizona Academic Content Standards	California Standardized Testing and Reporting Program	Massachusetts Comprehensive Assessment System	Virginia English Standards of Learning Curriculum Framework	Wisconsin Model Academic Standards
Norm-referenced test included	SAT-9			SAT-9	
Standards relating to reading components, subskills	▪ Strand 1: Reading Process	▪ Word analysis, fluency, and vocabulary development	▪ "Identify basic facts and main ideas in a text . . ."	▪ "Apply knowledge of word origins, derivations . . ."	▪ "Use effective reading strategies to achieve their purposes . . ."
Standards relating to reading contexts, passage types	▪ Strand 2: Comprehending Literary Text ▪ Strand 3: Comprehending Informational Text	▪ Reading comprehension (focus on informational materials) ▪ Literary response and analysis	"Identify, analyze, and apply knowledge of" ▪ different genres ▪ theme in literary ▪ structure and elements of fiction ▪ nonfiction and informational text ▪ elements of poetry ▪ an author's style and language ▪ myths, traditional narratives, and classical literature ▪ dramatic literature	▪ "Read and analyze a variety of narrative and poetic forms . . ." ▪ "Read, comprehend, and analyze a variety of informational sources . . ."	▪ "Read, interpret, and critically analyze literature." ▪ "Read . . .literary and nonliterary texts in order to understand human experience." ▪ "Read to acquire information."
Performance levels	▪ Approaches the standard ▪ Meets the standard ▪ Exceeds the standard	▪ Far below basic ▪ Below basic ▪ Basic ▪ Proficient ▪ Advanced	▪ Warning/failing ▪ Needs improvement ▪ Proficient ▪ Advanced	▪ Scored 0-600: ▪ 0-399 fail ▪ 400-600 pass ▪ (500-600 advanced)	▪ Minimal performance ▪ Basic ▪ Proficient ▪ Advanced

text characteristics, and it is difficult to interpret the results within the context of reading ability components as has been discussed here.

The relative difficulty of discerning reading components based on analyses of large-scale test scores suggests that the method itself may present difficulties to researchers. The results of other studies using similar methods indicate that expected relationships often do not emerge from analyses of large-scale test scores. Li, Ford, and Tompkins (1999) looked at the stability of content area scores between 3rd and 5th grade test takers of the Maryland School Performance Assessment Program, and found evidence that was more consistent with a general measure of student ability, rather than with specific content areas. Li (2001) looked at longitudinal true-score correlations between content areas of the Maryland test and the Comprehensive Test of Basic Skills (CTBS), and found more evidence to support the construct validity of the content areas, although more so for the multiple-choice CTBS test than for the Maryland test. In a study of the Washington state assessment (MacQuarrie, 2003), the author found moderate correlations between scores on the standards-based reading tests and norm-referenced tests given in the year prior, providing some criterion validity for the content standards.

Other studies have had difficulty finding psychometric evidence for hypothesized changes in test responses due to changes in reading ability. Perkins and Pohlmann (2002) looked at response patterns on an English as a Second Language (ESL) reading comprehension test administered three times to a subject pool whose English language competence was improving over that time period. The authors had expected to find evidence of growing competence in the changing patterns of responses to the tests, based on the theory that such growing competence results from a restructuring of knowledge structures. Instead, they found no psychometric evidence for such a restructuring, as the response patterns did not vary significantly from one test to the next.

Lack of Consensus on Reading Components

The available theoretical literature and applied models indicate that there is no consensus on whether reading should be viewed as a single ability or a multicomponent skill. Further, research indicates that when we go looking for components what we find (and how easily we find it) may be partially a function of the methods we use. For designers of large-scale reading tests, this divergence of opinion presents no clear guidance on how reading ability should be constructed. But it poses a particular challenge to the effort to base those tests on content standards that have been developed to meet a variety of needs. For if the research suggests that the number and quality of reading components that emerge from our analyses reflect to some degree the theoretical perspectives and assessment techniques we bring to the problem, then the development of standards-based test questions becomes a circular exercise in which both the standards and the questions depend on each other to be properly designed. Alderson's (1990) study raises the important distinction between the psychological validity and empirical validity of reading components. To argue that reading components exist is one matter; to demonstrate that they can be empirically isolated, measured, and thereby analyzed as variables, is quite another. Alderson (2000) surveys the empirical evidence for reading components, and finds that it is less than persuasive. Thus, he notes that the conservative position staked out many years ago by Lennon (1962) remains relevant to reading research today: “We still have little experimental evidence about the reality of the distinctions that are made among the various reading abilities and about the validity of supposed diagnostic profiles of reading skills” (Lennon, 1962, p. 332; cited in Alderson, 2000, p. 94).

Effect of Item Format on Test Performance

In the context of large-scale testing the relationship between the format of a test item and the ability or construct it is attempting to measure is an important aspect of validity. The formats that are generally applicable to discussions of large-scale tests are multiple-choice (MC), which asks the student to choose from among a small number (usually four or five) of alternatives, and constructed-response (CR), which ask the student to compose an answer. CR items can call for answers ranging from a single word, phrase, or number, to essays of several paragraphs, responses that have been labeled completion and construction, respectively (Traub, 1993). CR items are generally included in large-scale tests with the expectation that the CR format will elicit assessment information that cannot be obtained from MC items. Whether that expectation is warranted, and the usefulness of the assessment information gained by CR items, are of practical significance given the high costs of grading CR items.

The relationship between item format and test performance involves several related concerns. On the one hand, we would want to see evidence that both MC and CR formats elicit responses that are valid indicators of the same construct (for example, reading ability); that is, the formats are trait equivalent. This concern is often raised when considering the role of reading and writing abilities in answering CR items intended to measure a student's math ability. On the other hand, the formats need to elicit sufficiently different information in order to justify their inclusion in a test. In practical terms, this generally means that CR items should provide assessment information over and above that provided by MC items. We can also consider whether certain item formats elicit responses that misrepresent the construct of interest by requiring othe abilities or skills than that understood to be part of the intended construct. That is, does a format add variance that is irrelevant to the construct of interest? From a practical standpoint, then, Pearson and Garavaglia (2003) suggest that the key questions regarding item format are: "a) do constructed-response items provide us with more information about what students are capable of doing than we would get from multiple-choice items alone, and, if so, b) what types of skills are tapped by the constructed-response that are not measured by multiple-choice items?" (pp. 13-14).

Research into the notion that CR items contribute such information is inconclusive. Bennett (1993) states that previous empirical research has provided "only equivocal evidence" that MC and CR formats tap "fundamentally different" skills (p. 8). Traub (1993) argues that empirical evidence related to this issue tended to be specific to subject domains, and that for "reading comprehension and the quantitative domain, the answer is probably that tests that differ by format do *not* measure different characteristics" (p. 38, emphasis in original). Bridgeman and Rock (1993) examined the analytical reasoning test items of the GRE General test, in light of previous work that had indicated weak psychometric evidence for the construct validity of analytical reasoning as distinct from the verbal and quantitative abilities measured by other sections of the

GRE. The authors hypothesized that computer-based delivery of analytical reasoning items in a CR format would more accurately capture the construct of analytical reasoning than paper-based items had been able to, and thus indicate more support for the construct validity of analytical reasoning. However, the authors found that MC and CR item formats were highly correlated ($r = .93$), which led them to conclude that both MC and CR formats tapped the same construct, and that there were no discernable format effects.

Ercikan, Schwartz, Julian, Burket, Weber, and Link (1998) studied MC and CR items in order to determine whether their psychometric properties permitted them to be calibrated together within an item response theory (IRT) model framework. Of central importance was the question of whether both item types reflected the same underlying construct, and thus could be combined into one score, without sacrificing information about the test taker's performance. If the format of the items produced responses that diverged sufficiently to indicate that MC and CR items tapped different abilities, then combining the items would result not only in a loss of information but also violate the assumption of unidimensionality that is basic to IRT models. Ercikan et al. found no meaningful differences between item formats, and concluded that they could be combined to measure a single construct.

The effort to link item formats to different aspects of a construct or ability, such as reading, raises the issue of format effects on cognition. In an overview of research on the relationship between item format and cognition, Martinez (1999) noted that while both formats can accommodate a wide range of cognitive activity, "the range of cognitions within the reach of CR items is broader" (p. 209). Thus, he argues that CR items have the potential to elicit more complex thinking than do MC items, and consequently are more useful for diagnostic purposes. This potential is complicated, however, by the impact of construct-irrelevant factors such as the test taker's prior knowledge, answering strategies (such as reading test questions before the associated passage), test anxiety, and student expectations of format demands, all of which may affect MC and CR responses in different ways.

These complicating factors suggest that research methods have had a role in obscuring the differences between MC and CR items. Bennet (1993) pointed to several possible causes of the researchers' inability to detect these differences, all of which related to research design, including an over-reliance on correlational studies. Messick (1993) called for a less restrictive method of testing for trait equivalence that was more in line with construct validation methodologies. In a meta-analysis of 67 studies investigating the trait equivalence of MC and CR formats, Rodriguez (2003) found that such equivalence was largely a product of item design. The closer the two formats were to being essentially identical (i.e., stem equivalent) for a given construct, the more appropriate it was to combine their scores. Where the formats were used to tap different aspects of the construct or different cognitive abilities (as is the case with the HSA), combining the scores to reflect a single construct becomes more problematic.

Pearson and Garavaglia (2003) also concluded that the evidence does not enable researchers to attach item formats to "different characteristics or skills" (p. 15), but argue that the factor analytic methods that have often been used to answer these questions are problematic. They propose a construct-centered method of item format analysis that "begins with an explication of a theory of the domain being assessed," and then proceeds to articulate the specific aspects of that construct that can best be assessed with the different item formats (p. 18). While such an approach might avoid the weaknesses of factor analytic methods, however, it should be noted that explicating a theory of, for example, reading ability, is not a simple matter, as the discussion of reading components above indicates. Further, it seems unlikely that one could build consensus for such a theory in the absence of strong psychometric evidence from large-scale tests, which would bring the effort right back to the analytic methods that the authors find to be wanting.

It should be noted that the practical advantage that MC items possess for large-scale tests need not be viewed as defining a theoretical standard against which CR items are assessed. Lubliner and Smetana (2003) argue that the MC format elicits test-taking behavior that is inconsistent with the concept of reading comprehension. The authors found that their subjects' good performance on an MC test did not appear to help them with a CR test, although both were based on the same passages. Thus, they conclude that the MC format elicited test taking behavior that did not reflect engagement with or understanding of the text, which was evident in the poor performance on CR items. The authors argue that MC results reflect test-taking strategies that have nothing to do with actual text comprehension, and thus that MC items are an invalid measure of reading ability.

A strong position against the usefulness of MC items in assessing reading ability is presented in Katz and Lautenschlager's (2001) analysis of the reading comprehension task of the Scholastic Aptitude Test (SAT). Katz and Lautenschlager find that what they refer to as "no-passage factors"—that is, factors that enable students to answer reading comprehension questions without having read the passage on which those questions are based—account for far more systematic variance in performance on those questions than factors associated with the passage. The authors argue that their findings "reinforce the persistent doubts about the construct validity of a task whose ostensible purpose is to measure passage reading ability. If factors unrelated to the passages collectively are the dominant predictor of item performance on a passage-reading task, the obvious conclusion is that the task, despite face validity, has little to do [with] passage reading ability" (p. 173). Of the no-passage factors that influence performance on the reading comprehension items, the authors speculate that the most important is "outside knowledge," or prior knowledge of the subject of the reading passage, general knowledge "about the world," and knowledge of test-taking strategies. The authors conclude that the basic problem test developers face is that "we are presently very far from a full understanding of the reading process itself" (p. 174).

Katz and Lautenschlager's study contributes to a long-running debate over the validity of using MC items to assess reading ability that centers on the distinction between information provided by the MC test question itself (independent of the reading passage on which it is based), and the reading passage. Early studies found that subjects, when presented with MC questions only, were able to answer correctly at a better-than-chance rate. Findings like these indicated that what were called "item variables"—that is, information that could be obtained simply by reading the test questions—accounted for a greater proportion of MC reading test performance variance than should be expected, if one assumes that correct answers reflected competent reading of the passage on which those answers were based. The disproportional importance of item variables thus called into question the validity of the MC item as an accurate method of measuring reading ability; as Katz and Lautenschlager put it, “what precisely, then, does the [MC format] task measure?” (2001, p. 173). Rejoinders to what were called “extreme criticisms” of MC items argued that critics misclassified item variables, and that many of the factors attributed solely to MC questions were, in fact, attributes of the reading passage and hence more properly understood as “text variables” (Freedle & Kostin, 1994). On this basis, correctly distinguishing between item and text variables reduces performance variance attributable only to MC questions and restores the importance of reading the passage. Freedle and Kostin (1994) acknowledge, however, that their analysis “does not demonstrate that correct responses to multiple-choice items necessitate a coherent representation of the passage” (p. 110), and that further work using different methods (other than correlational studies) needed to be done to address this issue.

The issue of whether a specific item format can be used as an indicator of specific cognitive representations is addressed in a study by Ayala, Yin, Schultz, and Shavelson (2002). The authors attempted to determine if test items drawn from large-scale science achievement measures would fit into three reasoning dimensions that had emerged from previous analysis of the National Education Longitudinal Study of 1988. The authors examined both MC and CR formats, as well as performance assessments (PA) designed (and confirmed by expert review) to match each reasoning dimension. Results indicated that the CR items displayed poor reliability (and thus were eliminated from further analysis), and the remaining items did not converge on the dimensions as expected. On the basis of these findings, the authors suggest that the test taker's prior knowledge, rather than the format of any given test item, provides a better understanding of the test taker's reasoning while solving the item. This point is supported by Wang's (1999) study of the choices made by students taking the Advanced Placement Chemistry examination of CR items. Wang found that students chose items that appeared to be less difficult, a judgment that was based on the familiarity of the item subject matter. The test results indicated, however, that those items judged easier were, in fact, more difficult than items with less familiar content and chosen less often. Prior knowledge, in this case, played an important role is interpreting the results of CR test items.

Another reason that item format effects may not be clear is suggested by a study by O'Neil and Brown (1998), which looked at the impact of format on metacognitive and affective processes of students on a large-scale mathematics test. The authors found that open-ended questions generated more strategic thinking (metacognition) and worry (affect) than did MC questions. Item format effects on cognition, then, may reflect not only the demands of the format but also personal attributes of the test taker. Similarly, DeMars (2000) found that increasing the stakes of a test consisting of both MC and CR items generally resulted in better performance on both formats, but the performance increase on CR items was significantly larger than on MC items. DeMars speculated that item format may affect the motivation and performance of test takers in high stakes situations.

It appears, then, that item format effects on performance in a large-scale testing context tend to be inconsistently demonstrated in factor analytic and correlational studies. Advocates of CR items' ability to elicit assessment information over and above that obtained by MC items suggest that researchers need to adapt a more construct-centered approach to identifying and understanding how to use CR items to bring out that information. But there is also evidence to suggest that format differences, with respect to cognition, are entangled with a range of complicating factors such as students' prior knowledge, motivation, and emotions.

CHAPTER 3. RESEARCH METHOD

Relevant Standards for Assessing the HSA Reading Test

For this study, I have chosen to address four research questions on the basis of standards in the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999) that have direct relevance to the structure and scoring of the HSA reading test. The relevant standards are as follows:

1. Standard 1.10: "When interpretation of performance on specific items, or small subsets of items, is suggested, the rationale and relevant evidence in support of such interpretation should be provided" (p. 19)
2. Standard 1.11: "If the rationale for a test use or interpretation depends on premises about the relationships among parts of the test, evidence concerning the internal structure of the test should be provided" (p. 20)
3. Standard 1.12: "When interpretation of subscores, score differences, or profiles is suggested, the rational and relevant evidence in support of such interpretation should be provided" (p. 20)

Standard 1.11 can be applied to an analysis of the relationship between the aligned SAT-9 and DOE items, which contribute to a student's HSA score, versus that between the nonaligned SAT-9 and DOE items. The designation of test items as "aligned" versus "nonaligned" creates an explicit distinction between parts of the HSA Reading test that can be examined. Standard 1.11 can also be applied to the relationship between MC and CR item formats. CR items are included on the HSA Reading test, as they are on many other standards-based tests, because they are believed to provide assessment information that is qualitatively different than that provided by MC items. This premise can be examined with the test scores. Standards 1.10 and 1.12 can be applied to the assignment of test items to reading strands, and the reporting of scores for each of those strands. These standards can thus be applied to an examination of the construct of reading ability on which the test items are based. Standards 1.10 and 1.12 are also useful in approaching the question of whether passage types affect performance on the test items. Although passage type is not a distinction used in reporting the test results, the clusters of items are based on specific passage types. This clustering reflects the Range content standard of the HSA.

Research Questions and Methods

Source of Data Used for the Analyses

The data set as provided by the DOE included the results of the HSA test administered statewide in spring 2002 to students in grades 3, 5, 8, and 10. The data categories that were used for this study include: (a) scores for all items of the HSA Reading test, and (b) the proficiency level into which each student was placed, based on the HSA Reading score. The data provided by the DOE excluded records that are unusable (incomplete or invalidated tests) or inappropriate for inclusion in the analysis (due to special testing circumstances). The analyses discussed in this section are therefore based on 10,620 records for Grade 8 and 9,068 records for Grade 10.

Software Used for the Analyses

The SAS System version 8.2 (SAS Institute Inc., 1999-2001) was used for all of the statistical analyses described in this section, with the exception of the calculation of the attenuated correlation coefficients and Z-tests described in the section on the relationship between aligned and nonaligned SAT-9 items and DOE items. For that part of the study, the correlation and reliability coefficients were obtained with SAS, but the final calculations were done using Microsoft Excel for Windows 2003.

Relationship between SAT-9 Items and DOE Items

Do the correlations between the aligned SAT-9 items, the nonaligned SAT-9 items, and the DOE items, support the distinction between aligned and nonaligned SAT-9 items? This distinction is important because the aligned SAT-9 items are added to the DOE items to arrive at a student's total reading score, while the nonaligned SAT-9 items are excluded from the student's score. Thus, the designation of specific SAT-9 items as aligned with one of the reading strands has important consequences for a student's placement into a proficiency level and for interpretations of that student's performance on each of the reading strands. The rationale for designating specific SAT-9 items as aligned with one of the reading strands is based on the premise that such items share a common trait with the DOE items measuring that strand (Hawai'i Department of Education, 2003). It is reasonable to expect, then, that student performance on the DOE items would more closely match performance on the aligned SAT-9 items than on the nonaligned SAT-9 items. One would therefore expect a stronger direct correlation between scores on the aligned SAT-9 items and the DOE items, than that between the nonaligned SAT-9 items and the DOE items. Such a finding would provide criterion based validity evidence that the designation of aligned SAT-9 items is appropriate, and that the information obtained by the aligned SAT-9 items is different than that obtained by the nonaligned SAT-9 items.

To answer this research question, the following steps were taken:

1. The SAS PROC CORR procedure was used to determine the correlations between the aligned SAT-9 items and the DOE items, the nonaligned SAT-9 items and the DOE items, and between the aligned and nonaligned SAT-9 items.
2. The SAS PROC CORR procedure was used to determine the reliability (Cronbach's coefficient alpha) of the aligned SAT-9, nonaligned SAT-9, and DOE items.
3. The correlations obtained in step 1 were adjusted for the reliability coefficients obtained in step 2, and these adjusted correlation coefficients were then compared.
4. A Z-test was conducted to determine if the difference between the unadjusted correlations was significant. Specifically, the correlation between the aligned SAT-9 and DOE items was compared to the correlation between the nonaligned SAT-9 and DOE items, and the significance of this difference was determined. The significance test used is appropriate for situations like this one, where the correlations being compared include common items (in this case, the DOE items were common to both correlations) (Tabachnick & Fidell, 2001, pp. 145-147)

Components of Reading Ability

Do the HSA scores provide evidence that reading ability, as measured by the HSA reading test, comprises three reading components, or "strands," identified by the DOE? This analysis would provide construct-related evidence of the validity of the three strands upon which the DOE's standards of reading are based. For the HSA measurement of reading strands to be theoretically reasonable and diagnostically useful, the data should indicate that items comprising each strand behave differently from those in the other strands. The rationale for this approach can be stated in this way: If Comprehension Processes, Conventions and Skills, and Response represent actual components of reading competency, and if the HSA reading test items accurately measure those components, then the test scores should provide evidence that those components are distinguishable. Thus, students who are relatively weak in Comprehension Processes, for example, should perform less well on the items that are aligned with that strand than they would on the items aligned with the other two strands. Likewise, students who are strong in Conventions and Skills should perform better on those items than on those aligned with the other two strands. The responses to the test items should therefore tend to cluster in groups that approximately reflect the student's ability on the three reading strands. The assumption that the scores will cluster in this manner, and that the clusters can then be used to gauge student ability on the strands, is made explicit in the instructions given to teachers on how to interpret strand subscores as well as in the reporting narratives designed to provide constructive feedback to students and parents on how to interpret the subscores and improve performance on each strand (Hawai'i Department of Education, 2003).

As discussed earlier, the reading strands are conceptualized as being interrelated (Hawai'i Department of Education Office of Accountability and School Instructional Support/School Renewal Group, 1999), so one would not expect to see complete independence among the three strands; there will be readers at all levels who perform similarly on all the strands. Nevertheless, overall the responses should demonstrate that the strands reflect sufficiently discrete aspects of reading. Any such patterns of responses should be evident in a covariance or correlation matrix of the items, which displays the covariance/correlation between each item and every other item.

There are two models that can be reasonably inferred from the available source material regarding the theoretical structure of the reading strands. These models, illustrated in Figures 1 and 2 (for Grade 8) and Figures 3 and 4 (for Grade 10), are conceptually very similar, but the statistical analyses used to test them differ slightly. In the first model, designated as the one-tier model, the three strands are interrelated components of reading. Each component influences, and is influenced by, the others, and there is no hierarchical relationship among them; no single component determines the others. In the second model, designated as the two-tier model, the three strands are subcomponents of a general reading ability. The three strands are thus related to each other through this general ability, rather than directly with each other.

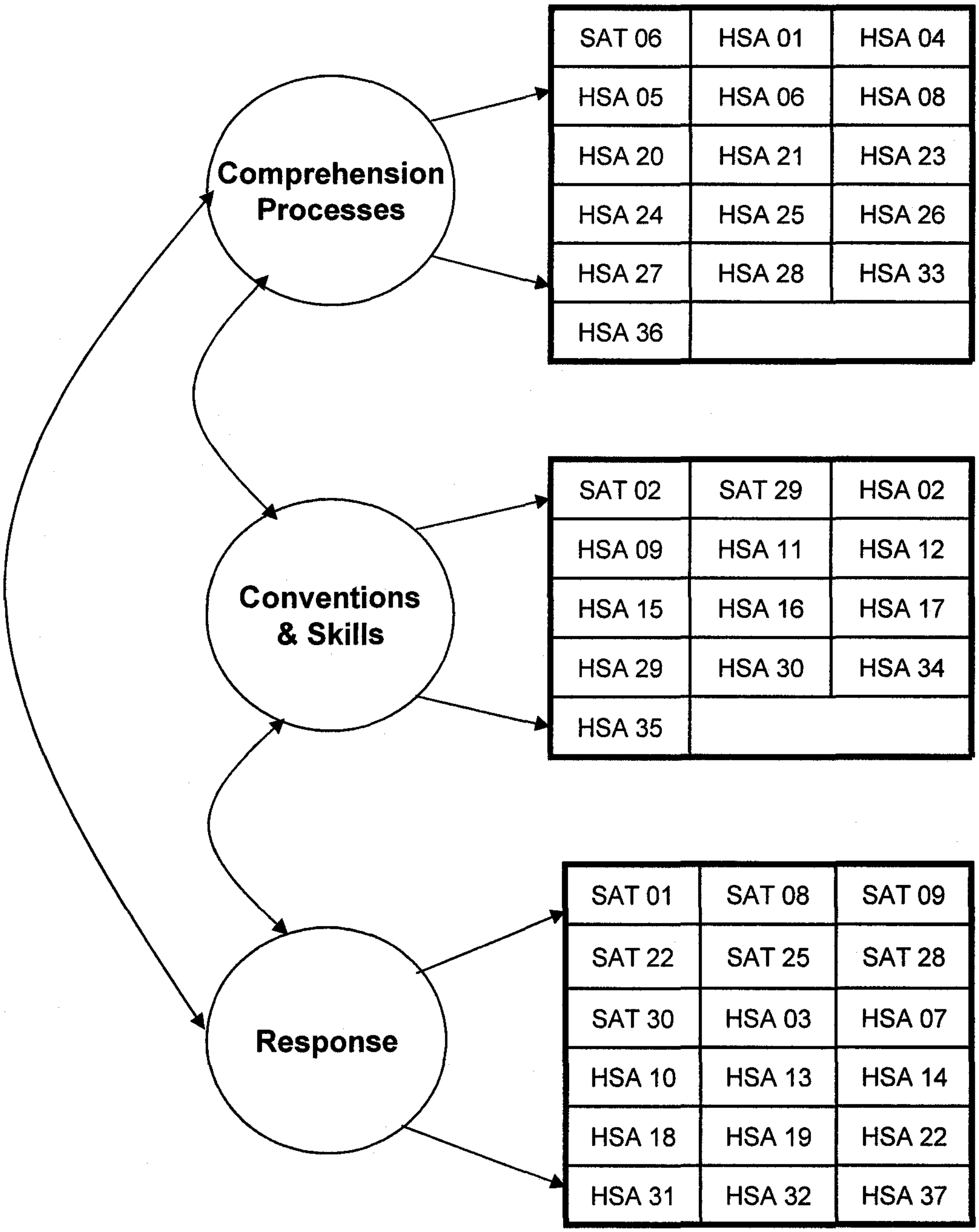


Figure 1. Grade 8 DOE Reading Model, One-Tier
(Note: For visual clarity, not all path lines to items shown)

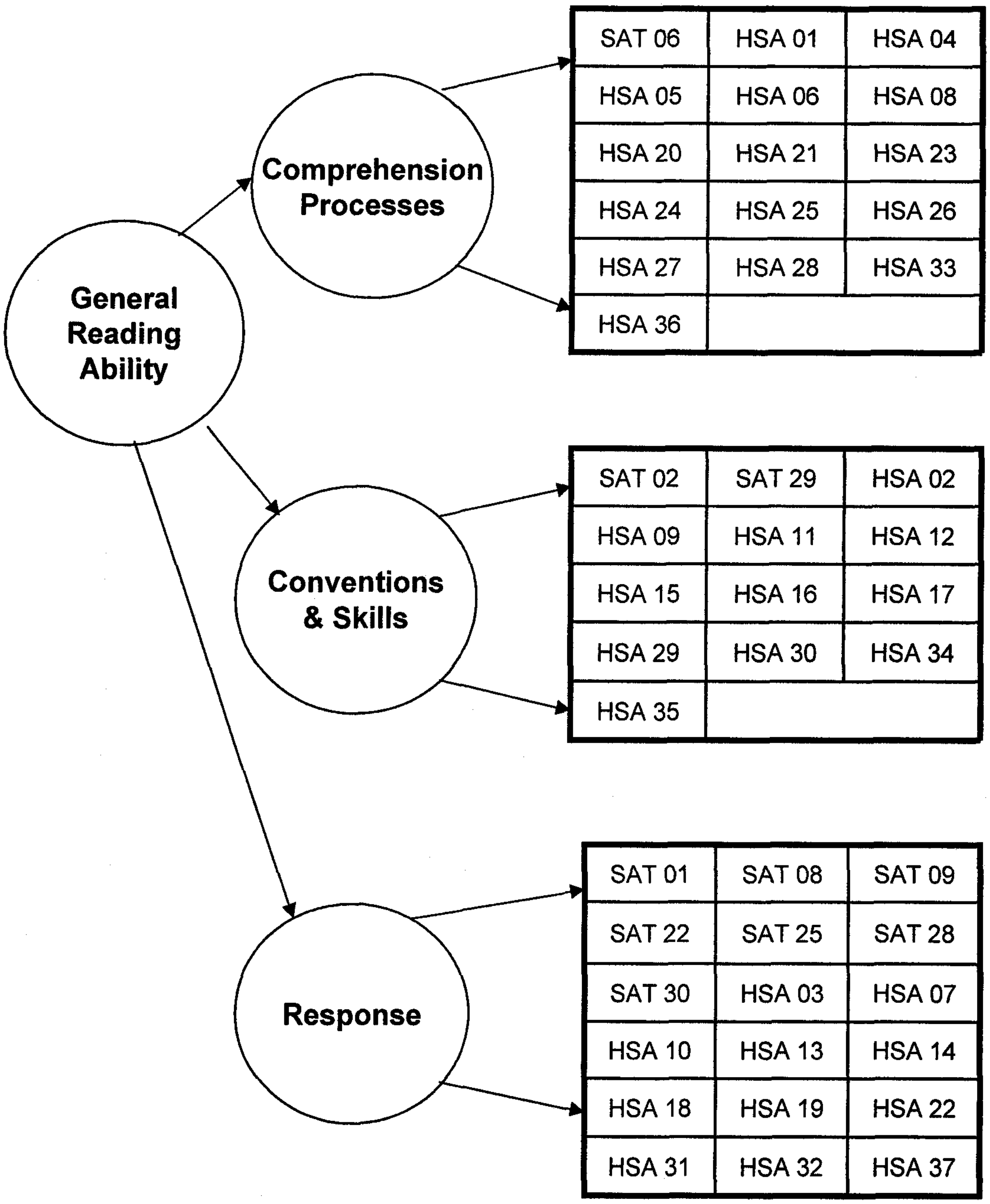


Figure 2. Grade 8 DOE Reading Model, Two-Tier
(Note: For visual clarity, not all path lines to items shown)

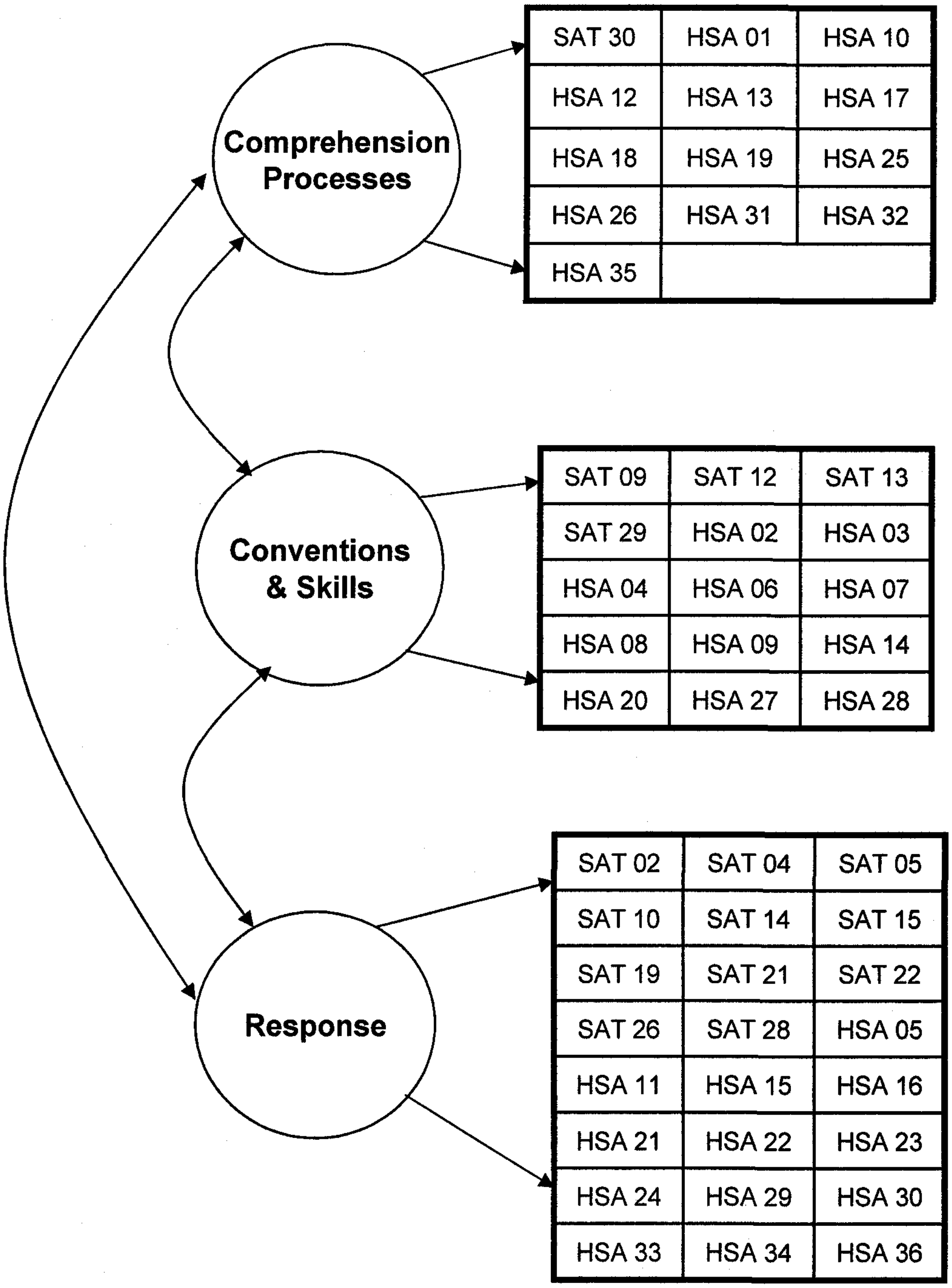


Figure 3. Grade 10 DOE Reading Model, One-Tier
(Note: For visual clarity, not all path lines to items shown)

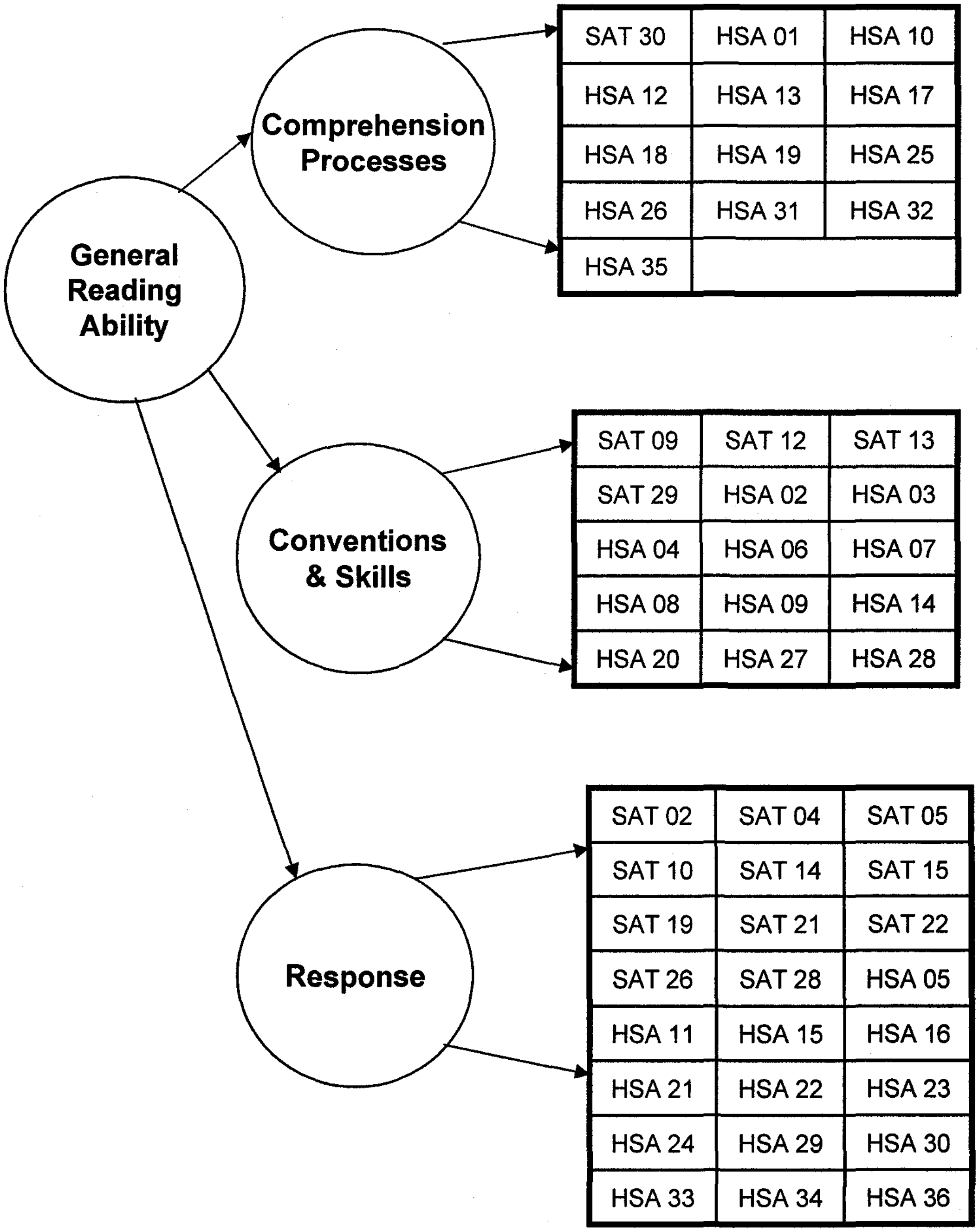


Figure 4. Grade 10 DOE Reading Model, Two-Tier
(Note: For visual clarity, not all path lines to items shown)

Research question 2 was addressed in the following steps:

1. Testing the DOE models: First, the one-tier and two-tier models of reading ability were tested against the data using a confirmatory factor analysis (for the one-tier model and the two-tier model) by evaluating the appropriate fit indices (discussed below) according to generally accepted fit criteria.
2. Assessing individual items: If the models did not meet the desired criteria, the loadings of individual items on their respective strands were examined for evidence of items that might be unduly affecting the indices due to very low loadings on the factor with which they were aligned. Any such items were then subjected to further tests to determine if there was sufficient evidence to warrant their exclusion from further analyses of model fit. If any such "bad" items were identified, the DOE models were tested again without those items.
3. Further testing of the models: After evaluating the fit of the DOE models based on the reduced data set (i.e., without the "bad" items), other possible problems with the models were considered. In particular, the high inter-factor correlations (in the one-tier model) and high loadings of each strand on the general reading factor (in the two-tier model) that emerged from the tests in steps 1 and 2 were further explored for possible evidence of item unidimensionality. Analyses were then conducted to see if a single-

component (unidimensional) model of reading ability better fit the data. This effort proceeded in three basic stages:

a. First, three alternative models were compared to see which one, if any, best fit the data: the two original DOE models (one-tier and two-tier), and a model with a single component (one general reading factor).

b. Subsequently, a set of "random assignment" models was created for testing. For each of these models, test items were randomly assigned to the three strands by creating a random sequence of the items, which were then simply distributed in that order to the three strands while retaining the original number of items per strand. These three models were tested twice: once within the one-tier structure, and again within the two-tier structure.

c. Finally, a "rotated assignment" model was tested within both the one-tier and two-tier structures. This model was created by successively assigning each item to each of the three strands. Thus, item 1 was assigned to the Comprehension Processes strand, item 2 to the Conventions and Skills strand, item 3 to the Response strand, item 4 to the Comprehension Processes strand again, and so on.

Confirmatory factor analysis produces indices that are used to gauge the degree to which the model being tested "fits" the actual data. The term "fit" generally refers to the comparison between the covariance/correlation matrix of the actual data and the covariance/correlation matrix produced by the model being tested. The closer the two matrices are to each other (that is, the smaller the differences between them), the better the fit. There are numerous fit indices; the SAS PROC CALIS procedure used in this study generates 23 such indices.

There is no agreement, however, on which fit indices are best to use (Maruyama, 1998; Thompson, 2000). Furthermore, for many indices the values that indicate good model fit are imprecise, and researchers rely on "rules of thumb" in the absence of definitive guidelines. Faced with a surplus of fit indices and a lack of consensus on which ones are optimal in given circumstances, researchers have offered classification schemes that provide some guidance on the appropriate use of indices (Byrne, 1998; Loehlin, 2004; Maruyama, 1998). Based on reviews and classifications of SEM fit indices, for the purpose of testing models of reading ability components in this study the following indices were chosen.

1. The comparative fit index (CFI), which is a measure of fit that adjusts for, and is applicable to, samples of any size (Hatcher, 1994). The CFI ranges from zero (indicating no fit) to one (perfect fit), with values close to 1.00 generally indicating adequate fit.
2. The non-normed fit index (NNFI) compares the tested model to a hypothetical model of no fit, but it also accounts for the complexity (or, conversely, the parsimony) of the tested model (in terms of the number of

parameters) (Bentler & Bonett, 1980, cited in Maruyama, 1998). Values above 0.90 generally indicate adequate fit.

3. The chi-square (χ^2) statistic and the degrees of freedom (*df*) associated with it are included as indications of model fit. A χ^2/df ratio equal to 1.00 would indicate perfect model fit; ratios of 2.0-2.5 indicate acceptable fit (Hatcher, 1994). The χ^2 significance test was not used to assess model fit in this study. The χ^2 significance test has been shown to be very sensitive to sample size, and when used with large samples it leads to unwarranted rejection of the null hypothesis; specifically, it leads to rejection of the tested model even when its fit is acceptable under other criteria (Thompson, 2000). Thus, it is an inappropriate indicator of fit for the analyses in this study, which were based on populations of 10,620 and 9,068 students in Grades 8 and 10, respectively. Given such large numbers of subjects (N), it was expected that the χ^2 significance test would lead to rejection of all tested models. This is, in fact, what happened; all of the χ^2 significance test results rejected the tested models at $p < .0001$.

Contribution of Constructed-Response Items

Do the open-ended (OE) and extended-response (EX) items, which make up the constructed-response (CR) format items on the test, contribute information to a student's placement into one of the four proficiency level categories, over and above the information provided by the multiple-choice (MC) items? This analysis would provide evidence that constructed-response items elicit responses that contain assessment information that is not obtained only by multiple-choice items. As can be seen in Table 4, multiple-choice items account for a greater percentage of items and possible points than do the constructed response items. Yet, the fact that constructed-response items are included, in spite of the greater complications and costs associated with grading them, reflects the importance that advocates attach to the quality of assessment information the CR format provides.

Table 4

Score Structure of the HSA Reading Test by Response Format

	Multiple Choice	Open Ended	Extended	Total
Grade 8				
Number of Items	39	6	2	47
Point Value	1	3	4	
Total Possible Points	39	18	8	65
Maximum Possible (%) Contribution to Total Score	60.0	27.7	12.3	100.0
Grade 10				
Number of Items	44	5	3	52
Point Value	1	3	4	
Total Possible Points	44	15	12	71
Maximum Possible (%) Contribution to Total Score	62.0	21.1	16.9	100.0

Discriminant analysis (DA) was used to assess the contribution of each response format's scores to student placement into a proficiency level. DA refers to a set of multivariate procedures concerned with the classification of subjects into one of several categories (or groups) of a categorical variable, based on functions or equations derived from those subjects' scores on other variables. DA procedures enable researchers to address various aspects of classification, including determining the best variables to classify subjects, assessing the accuracy of the equations used to classify subjects, and predicting the classification of subjects (Duarte Silva & Stam, 1995; Stevens, 1992), and have been used to detect the best predictors of student placement in standards-based performance levels (Good, 2002).

DA was used to examine the contributions of each of the three response formats (MC, OE, and EX) to the classification of students into one of the four proficiency level categories (below, approaching, meeting, and exceeding standards) in two basic steps. First, a stepwise analysis was used to determine which of the response formats contributed significant information to the placement of students into a proficiency level class. The formats identified as significant in this procedure were then used in a second analysis, in which the incremental improvement in correctly predicting proficiency levels was calculated for OE and EX item formats.

Relationship Between Passage Type and Test Performance

Is performance on literary, informational, and functional passage types related to a student's overall performance on the test? As noted earlier, the inclusion of these three passage types reflects the DOE's language arts strand of Range, which states that students should be able to comprehend a variety of different text types. The relationship between passage type and overall test performance was explored through repeated-measures analysis of variance (ANOVA) procedures. The primary variable of interest—the performance of students on each type of passage—was defined as the percent of total possible points earned by a student for each of the three passage types. Thus, each student was measured on each of the passage types by dividing his or her score by the maximum number of points possible on each of the passage types.

The first step of the analysis was to address the question of whether there were differences in student performance among the three passage types in general. For this question, a one-way ANOVA with repeated measures (using the SAS PROC GLM procedure) was conducted to test the null hypothesis that students performed equally well on all three passage types. In addition, contrasts were done between each pair of passage type to determine which pairs were significantly different. In the second step of the analysis, a two-way ANOVA with repeated measure on the independent variable of passage type was used to detect the interaction, if any, between passage type and proficiency level. Finally, a one-way ANOVA with repeated measure was used to examine the pattern of performance on the three passages types within each proficiency level category.

CHAPTER 4. RESULTS

Relationship between SAT-9 Items and DOE Items

The adjusted and unadjusted correlations shown in Table 5 demonstrate that the relationship between the SAT-9 and DOE items differed between the two grade levels. As noted above, we would expect a higher correlation between the aligned SAT-9 and DOE items than between the nonaligned SAT-9 and DOE items. For Grade 8, the results ran contrary to that expectation; the correlation between the aligned SAT-9 and DOE items was lower than that between the nonaligned SAT-9 and DOE items. The Grade 10 results, on the other hand, followed the expected pattern, with the aligned SAT-9 items showing a higher correlation with the DOE items than did the nonaligned SAT-9 items. For both grades, the difference between the correlations was significant ($Z = -9.89345$, $p < .05$ for grade 8, $Z = 6.413292$, $p < .05$ for Grade 10). However, the significant Z test results should be interpreted with caution, given the large number of subjects on which the tests were based and the small effect sizes between the correlations. A conservative interpretation of the results is that there was little difference between the correlations; that is, the aligned and nonaligned SAT-9 items were essentially equivalent in terms of their shared variance with the DOE items.

The reliability coefficients (Cronbach's alpha) on which the adjusted correlations were based, shown in Table 6, indicate that both the DOE and SAT-9 segments appear to have been consistent measures of the same reading ability. Thus, for both grades the reliability of the HSA score segment (aligned SAT-9 and DOE items) was higher than

Table 5

Adjusted and Unadjusted Correlations Between SAT-9 and DOE Questions

	Aligned SAT-9	Nonaligned SAT-9	DOE
Grade 8			
Aligned SAT-9	-----	-----	-----
Nonaligned SAT-9	1.00 (.68)	-----	-----
DOE	.82 (.61)	.87 (.68)	-----
Grade 10			
Aligned SAT-9	-----	-----	-----
Nonaligned SAT-9	1.00 (.67)	-----	-----
DOE	.85 (.65)	.79 (.60)	-----

Table 6

Reliability (Cronbach's Alpha) Estimates for Test Segments

	All Items	HSA Score	All SAT-9	DOE	Nonaligned SAT-9	Aligned SAT-9
Grade 8	.90	.88	.81	.86	.71	.65
Grade 10	.91	.89	.81	.87	.67	.68

either the SAT-9 or DOE segments. Further, the coefficients indicated that including the nonaligned SAT-9 items in the HSA score would have improved the reliability of that score.

The correlations between the SAT-9 items and DOE items are similar to correlations between local and national tests in other jurisdictions. For example, the correlations between the reading scores on the standards-based Alaska State Student Assessment and those on the California Achievement Test Fifth Edition in the spring 2000 test administration were .78 for both grades 8 and 10 (Fenton, 2003). Like the HSA, the Alaska State Student Assessment includes both norm-referenced and standards-based items in order to generate both types of information from the test results. Similarly, the correlation between reading scores on the standards-based Illinois Standard Achievement Test and the norm-referenced Iowa Test of Basic Skills in 2002 was .85 for Grade 8 (Easton et al., 2003).

Components of Reading Ability

As discussed in the chapter on methods above, tests of a three-component reading model proceeded in several steps. The results of all of the tested models for grades 8 and 10 are presented in Table 7.

1. With respect to the DOE models, for both grades the CFI and NNFI fit indices indicated reasonable fit; they approached, but did not reach, the .90 threshold that is generally used to indicate adequate model fit. The χ^2/df ratios suggested relatively poor fit based on "rule-of-thumb" guidelines, although as noted above this is expected from any index based on the chi-square statistic, which is strongly influenced by sample size.

Table 7

Fit Indices and Inter-Factor Correlations for Reading Models, Grades 8 and 10

One-Tier (Confirmatory Factor Analysis)

Grade 8			Random Assignment			Rotated
Fit Index	DOE	1 Factor	1	2	3	Assignment
CFI	0.86	0.85	0.85	0.85	0.85	0.85
NNFI	0.85	0.84	0.84	0.84	0.85	0.84
χ^2/df	12.72	13.24	12.96	13.11	12.79	13.11
Inter-factor correlations						
$r_{\text{comp-conv}}$	0.97		0.94	0.95	1.00	1.00
$r_{\text{comp-resp}}$	0.92		0.95	1.00	0.96	0.97
$r_{\text{conv-resp}}$	0.91		1.00	0.95	0.90	1.00
Grade 10			Random Assignment			Rotated
Fit Index	DOE	1 Factor	1	2	3	Assignment
CFI	0.81	0.81	0.81	0.83	0.83	0.83
NNFI	0.82	0.82	0.80	0.82	0.82	0.82
χ^2/df	10.30	10.42	11.46	10.42	10.40	10.31
Inter-factor correlations						
$r_{\text{comp-conv}}$	0.98		1.00	1.00	1.00	0.96
$r_{\text{comp-resp}}$	1.00		0.97	1.00	1.00	1.00
$r_{\text{conv-resp}}$	0.93		0.94	1.00	1.00	0.95

Two-Tier (Structural Equation Model)

		Random Assignment			Rotated
Fit index	DOE	1	2	3	Assignment
CFI	0.86	0.85	0.85	0.85	0.85
NNFI	0.85	0.84	0.84	0.84	0.84
χ^2/df	12.73	13.26	13.15	13.25	13.21
Std. Loadings on general reading factor					
Comp	0.99	0.99	0.99	1.00	1.00
Conv	0.98	1.00	0.97	1.00	1.00
Resp	0.93	1.00	0.97	1.00	0.97
		Random Assignment			Rotated
Fit index	DOE	1	2	3	Assignment
CFI	0.83	0.82	0.83	0.83	0.83
NNFI	0.82	0.81	0.82	0.82	0.82
χ^2/df	10.32	10.88	10.43	10.43	10.32
Std. Loadings on general reading factor					
Comp	1.00	1.00	1.00	1.00	1.00
Conv	0.95	1.00	1.00	1.00	0.95
Resp	0.98	0.95	1.00	1.00	1.00

Note: Abbreviations for reading strands:
Comp = Comprehension Processes
Conv = Conventions and Skills
Resp = Response

2. Subsequent examination of the loadings of each item on its respective strand suggested that four items in the Grade 8 test that might be disproportionately weakening model fit: SAT question 8, and HSA questions 22, 28, and 34. No such items were indicated by the loadings for the Grade 10 test. The low loadings of these four Grade 8 items ranged from .01 to .15, as shown in Table 8, which summarizes the evidence for identifying the four items in the Grade 8 test. In an exploratory factor analysis specifying a three-factor solution, all four items loaded poorly on the first factor, which accounted for 89 percent of the common variance, as well as on the subsequent two factors, which accounted for the remaining 10 percent of the common variance. The final column of Table 8 demonstrates that each of the four items was poorly correlated with the other items of the HSA Reading test.

Table 8

Evidence of "Bad" Items in the Grade 8 Test

Item	Standardized Loadings on Aligned Factor, DOE Model		Loading on First 3 Common Factors			Correlation with Other HSA Score Items
	One-Tier	Two-Tier	1st	2nd	3rd	
SAT 8	.15	.15	.16	.02	.07	.15
HSA 22	.15	.15	.15	.02	-.04	.14
HSA 28	.01	.01	.01	-.05	.07	.00
HSA 34	.04	.04	.04	.00	.08	.03

3. Parsing these four “bad” items from the Grade 8 model tests resulted in positive, but negligible, improvements (ranging from .002 to .003) to the CFI and NNFI indices; thus, the fit of the DOE models remained moderate for both grades. Attention was then directed towards the high inter-factor correlations among the strands in the one-tier DOE model (ranging from .91 to .97), and the high loadings between each strand and the general reading factor in the two-tier model (ranging from .93 to .99) that emerged from the fit tests, and which strongly suggested that the reasonable fit indices might have resulted from shared variance among the strands. This was confirmed by the tests of the random and rotated assignment models, as well as a single-factor model (for the one-tier model), all of which yielded fit indices and inter-factor correlations that were virtually identical to those of the DOE models.

The substantial equivalence of the tested models, and the lack of sensitivity of the factors to the items that comprised them, is apparent by examining the narrow range of values for each fit index. Across all models tested for Grade 8, the CFI ranged from .85-.86, the NNFI ranged from .84-.85, and the χ^2/df ratio ranged from 12.72 to 13.26. For Grade 10, the CFI ranged from .81 to .83, the NNFI ranged from .80 to .82, and the χ^2/df ratio ranged from 10.30 to 11.46. The substantial equivalence of the models is also evident in the inter-factor correlations among the three strands in the one-tier models, and between each of the strands and the

> hypothesized general reading factor in the two-tier models. For Grade 8, the inter-factor correlations among the three strands in the one-tier models ranged from .91 (between the Conventions and Skills and Response strands as originally composed) to 1.00; for the two-tier models, the standardized loadings of the strands on the general reading factor ranged from .93 to 1.00. For Grade 10, the inter-factor correlations among the three strands in the one-tier models ranged from .93 (again, between the Conventions and Skills and Response strands as originally composed) to 1.00; for the two-tier models, the standardized loadings of the strands on the general reading factor ranged from .95 to 1.00. These inter-correlations and loadings indicated each of the strands shared at least 80 percent of its variance with the others.

Overall, the results indicated that whether reading ability was operationalized as a single factor, as three inter-related factors, or as three factors subsumed under an overarching single factor, the HSA reading data support these models equally. Furthermore, the substantial equivalence of the eight alternative models in which items were randomly assigned to the three factors (three randomized plus one rotated assignment model for both the one-tier and two-tier structural models) indicated that strength of the fit indices did not depend on the proper alignment of test items with strands. Instead, the results indicated that models containing arbitrary assignments of test items to strands are comparable, if not identical, to each other and to the DOE models in terms of their fit of the data.

Contribution of Constructed-Response Items

Before conducting the discriminant analysis (DA), reliability coefficients (Cronbach's alpha) for each response format were calculated. As expected, given the different numbers of MC, OE, and EX items, higher reliability coefficients were found for the MC format. For Grade 8, the coefficients were: .85 for MC items, .72 for OE items, and .54 for EX items. For Grade 10, the coefficients were: .86 for MC items, .41 for OE items, and .40 for EX items. Thus, it appears that although response consistency was high and almost equivalent across both grades, both types of constructed-response items were less reliable on the Grade 10 test.

For the first part of the analysis for this research question, the SAS PROC STEPDISC procedure was used to conduct a stepwise discriminant analysis. In this procedure, a forward selection method begins with no variables in the model being used to explain the separation of subjects into groups, and then adds one variable at a time in descending order of explanatory power, provided that each added variable meets the minimum level of explanatory power specified by the researcher. After each variable is selected in this manner, all of the variables that have been added to the model to that point are evaluated according to their explanatory power, and any that do not meet the specified criterion are dropped from the model. This selection and retention process stops after a predetermined number of variables have been selected, or when none of the unselected variables meets the criterion for entering the model. For both the selection and retention of variables, the researcher selects the method and threshold for these actions. The method that was used in this analysis to evaluate variables—the MC, OE, and EX formats—for selection and retention in the model was the significance level of the *F*-test of the variable under consideration. The significance level chosen was $p < .05$, which is more conservative than the default value of .15 used by the SAS PROC STEPDISC procedure (SAS Institute Inc., 1999).

For both grades, all response formats were selected and retained in the model explaining student classification into proficiency level groups. For Grade 8, the values were as follows: MC format, $F = 8958.26$, $p < .0001$; OE format, $F = 1307.19$, $p < .0001$; and EX format, $F = 157.27$, $p < .0001$. For Grade 10, the values were: MC format, $F = 7697.04$, $p < .0001$; OE format, $F = 822.48$, $p < .0001$; and EX format, $F = 222.75$, $p < .0001$.

The second part of the analysis used SAS PROC DISCRIM to derive linear combinations of student scores on each of the item formats that most accurately place students into proficiency levels, and to assess the accuracy of those linear functions in predicting placement. The relative contribution of each format to placement accuracy was determined by comparing error rates (the rate of incorrect placement into proficiency levels). While there are different methods used to assess placement accuracy, the method used here was based on the simple number of placement errors. The baseline rates of correct placement in each level are set equal to the proportional number of cases in each of the levels in the total sample. The leave-one-out estimation method of predicting placement was used. In this method, each subject's placement is predicted by "leaving it out" and predicting its placement on the basis of the data of the remaining subjects (Duarte Silva & Stam, 1995). Table 9 shows the error rates for three linear discriminant functions for each grade level. The first function included only the MC items, the second included MC and OE items, and the third included all item formats. For both grades, MC items alone predict placement fairly accurately, with an error rate of 13-14 percent. The inclusion of OE items reduces that error rate by approximately one-half, and the full model again reduces the error rate by about one-half. Thus, the inclusion of all items in the prediction models results in a correct placement rate of 96 percent for both grades.

Table 9

Error Rates for Linear Discriminant Functions

Variables in Function	Error Rate	Improvement Over Previous Model
Grade 8		
MC	.14	
MC, OE	.07	.07
MC, OE, EX	.04	.03
Grade 10		
MC	.13	
MC, OE	.07	.06
MC, OE, EX	.04	.03

Relationship Between Passage Type and Test Performance

Initial tests were performed to determine whether students performed equally well on all three passage types; performance was measured by the mean percentage correct (based on scores) on all items belonging to each passage type. Table 10 includes the results of one-way ANOVA with repeated measure tests of performance across all passage types, which resulted in rejection of the null hypothesis of equal means in each grade level. Table 10 also includes the results of subsequent contrasts between each pair of passage type, with the direction of difference indicated. These results indicate that Grade 8 students generally did best on informational passages, then literary passages, and then most poorly on functional passages. Grade 10 students also did best on informational passages, but this was followed by middle performance on functional passages, and poorest performance on literary passages.

Table 10

ANOVA Summary: One-Way with Repeated Measure on Passage Type

Source	*df*	*SS*	*MS*	*F*
Grade 8				
Passage Type	2	32.42	18.21	1620.17*
Informational > Literary	1	2.49	2.49	104.75*
Literary > Functional	1	36.61	36.61	3028.45*
Informational > Functional	1	58.17	58.17	2402.27*
Grade 10				
Passage Type	2	7.82	3.91	354.62*
Informational > Literary	1	13.69	13.69	714.59*
Functional > Literary	1	9.39	9.39	311.41*
Informational > Functional	1	.40	.40	23.91*

* $p < .0001$

Subsequently, a two-way ANOVA with repeated measure was performed to assess the effects of passage type and overall reading skill, as measured by the student's proficiency level. Main effects were found to be significant for both passage type and proficiency level. Further, an interaction effect between passage type and proficiency level was found to be significant for both Grade 8 and Grade 10. These results are shown in Table 11.

The nature of this interaction is clarified by the results of one-way ANOVA with repeated measure tests, in which the simple effect of passage type was examined for students within each proficiency level. Mean percentage correct scores across the three

Table 11

ANOVA Summary: Two-Way with Repeated Measure on Passage Type

Source	*df*	*SS*	*MS*	*F*
Grade 8				
Between Subjects				
Proficiency Level	3	573.17	191.06	12292.60*
Error	10616	165.00	.02	
Within Subjects				
Passage Type	2	4.66	2.33	238.21*
Level X Passage	6	4.65	.78	79.23*
Error	21232	207.84	.01	
Grade 10				
Between Subjects				
Proficiency Level	3	524.64	174.89	10062.70*
Error	9064	157.52	.02	
Within Subjects				
Passage Type	2	.43	.21	19.60*
Level X Passage	6	2.25	.37	34.32*
Error	18128	197.82	.01	

passage types within each proficiency level are displayed in Table 12. A summary of the significance of the differences among those means is displayed in Table 13. The nature of the interaction varies between the two grade levels. For Grade 8, students in the lower two proficiency level categories perform best on literary passage items, second best on informational passage items, and worst on functional passage items. For students in the higher two proficiency level categories, this pattern changes. These students perform best on informational items, second best on literary items, and worst on functional items.

Table 12

Mean Passage Performance (Percent Correct) by Proficiency Level

Proficiency Level	N	Literary	Informational	Functional
Grade 8				
Below	596	.28	.26	.22
Approaches	4912	.52	.50	.44
Meets	4917	.70	.74	.65
Exceeds	195	.87	.90	.86
Grade 10				
Below	597	.23	.22	.21
Approaches	4327	.39	.44	.44
Meets	4019	.63	.66	.65
Exceeds	125	.86	.88	.87

For Grade 10, differences in performance within groups are generally smaller than for Grade 8, which is reflected in the relatively fewer significant values in Table 13. Students in the Below Proficiency group follow the same pattern as the Grade 8 students in the lowest two groups, but the differences between the means are very small (.02 between the highest and lowest means). Students in the Approaches Proficiency group show a distinctive pattern, doing equally well on informational and functional items, and less well on literary items, with the range of .05 between highest and lowest being the largest such range among all groups. Students in the Meeting Proficiency and Exceeding Proficiency groups show different patterns, but again the means are very close, suggesting that for them, as for students in the Below Proficiency group, there is no real difference in performance across passage types.

Table 13

Significance of Differences between Passage Types within Proficiency Levels

Proficiency level	All Means	Literary and Informational	Literary and Functional	Informational and Functional
Grade 8				
Below	$F(2, 1190)=54.97^{**}$	$F(1,595)=5.15$	$F(1,595)=154.6^{**}$	$F(1,595)=48.65^{**}$
Approaches	$F(2,9822)=737.25^{**}$	$F(1,4911)=25.58^{**}$	$F(1,4911)=2178.24^{**}$	$F(1,4911)=733.46^{**}$
Meets	$F(2,9832)=1143.43^{**}$	$F(1,4916)=531.79^{**}$	$F(1,4916)=850.71^{**}$	$F(1,4916)=1878.82^{**}$
Exceeds	$F(2,388)=25.18^{**}$	$F(1,194)=17.90^{**}$	$F(1,194)=9.51^{*}$	$F(1,194)=40.81^{**}$
Grade 10				
Below	$F(2,1192)=6.40^{*}$	$F(1,596)=1.85$	$F(1,596)=9.07^{*}$	$F(1,596)=6.54^{*}$
Approaches	$F(2,8652)=357.17^{**}$	$F(1,4326)=695.76^{**}$	$F(1,4326)=361.05^{**}$	$F(1,4326)=4.74$
Meets	$F(2,8036)=82.96^{**}$	$F(1,4018)=172.42^{**}$	$F(1,4018)=58.93^{**}$	$F(1,4018)=16.80^{**}$
Exceeds	$F(2,248)=1.34$	$F(1,124)=3.12$	$F(1,124)=.71$	$F(1,124)=.52$

$^{*}p < .0167$. $^{**}p < .0001$

Review of Findings

The expected stronger direct relationship between aligned SAT-9 and DOE items versus nonaligned SAT-9 and DOE items is supported only by the Grade 10 data. The Grade 8 data indicate the opposite effect, with the nonaligned SAT-9 items showing a stronger direct correlation with the DOE items. However, for both grades the effect sizes are small enough to suggest that, in practical terms, the correlations could be considered equal. Thus, both aligned and nonaligned SAT-9 items function similarly.

Latent factor analyses show that, for both grades, the specific three-strand model embodied in the HSA Reading test fit the data with only moderate success. Further tests indicate that random three-factor models and single-factor models produce virtually identical fit indices. Correlations among the strands indicate substantial shared variance among them. Given these results, the single-factor model is the most parsimonious solution. Thus, the data support a single-component or unidimensional model of reading.

Discriminant function analyses show support for the uniqueness of information provided by both open-ended and extended-response item formats, over and above that provided by multiple-choice items. For both grades, then, the evidence indicates that multiple-choice and constructed-response items elicit information that is useful in placing students into performance levels. However, the amount of information provided by constructed response items (in terms of accurately placing students into proficiency levels) is very small relative to that provided by multiple-choice items, even though the OE and EX items contributed 40 and 38 percent of the possible total score at Grades 8 and 10, respectively. The relatively low accuracy in placing students may be due to the insufficient internal consistencies of the OE and EX response formats.

Finally, ANOVA analyses indicate that the difficulty of different passage types depends on the proficiency level of the student. This interaction effect is stronger in Grade 8, with literary passage items easiest for students below and approaching proficiency, but informational passage items easiest for students meeting and exceeding proficiency. The Grade 10 results also indicate interaction between passage type and proficiency level, with a pattern that contrasts that of Grade 8, with literary passage items appearing to be the most difficult.

CHAPTER 5. DISCUSSION, IMPLICATIONS, CONCLUSION

Discussion of Findings

Alignment and Reading Strands

In light of the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999) discussed in Chapter 3, the findings concerning aligned and nonaligned SAT-9 items and reading strands (research questions 1 and 2) have significant implications for the interpretation and use of the HSA test scores. Specifically, although the data indicate that the DOE items have high reliability, they also indicate that the DOE items are highly correlated with all SAT-9 items, both aligned and nonaligned. Further, the data do not support the three-component model of reading upon which the specification of SAT-9 items to be included in a student's score is based, and upon which the division of a student's total reading score into subscores indicating performance on each of the strands is based. Based on these results, we can draw the following implications:

1. All SAT-9 and DOE items appear to measure the same ability with high reliability. Thus, the SAT-9 items have high predictive validity with respect to the student's performance on the DOE items, and vice versa. Since all SAT-9 items function in essentially the same way, there appears to be no rationale for distinguishing between aligned and nonaligned items. In fact, the results of correlational analyses indicate that including all SAT-9 items (rather than only the aligned items) in the HSA score improves the overall reliability of the test.

2. The results do not support the original alignment of the test items with the three reading strands. Rather, the results suggest that a unitary concept of reading ability is a more parsimonious and appropriate representation of the construct that the HSA Reading test is measuring. The lack of clear evidence in support of the DOE reading strands model of reading ability suggests that any interpretation of a student's reading ability beyond his or her total score must be approached with great caution. In particular, the use of strand subscores for diagnostic purposes, as indicated by the reporting narratives designed to explain to teachers, parents, and students what a low subscore in each strand means (Hawai'i Department of Education, 2003) and how the student can improve performance on that strand, is not supported by the evidence here.

Item Format and Passage Type

Findings regarding item format and passage type suggest the following:

1. The constructed-response items provide unique assessment information for the placement of students into proficiency levels, but the amount of that information is minimal relative to that provided by multiple-choice items. Thus, the inclusion of constructed-response items needs to be weighed against the time and resource costs required to grade those items.
2. The relatively low internal consistency that characterizes the constructed-response items, particularly the EX items, suggests that these items may have difficulty functioning as effectively in large-scale test situations as

intended. Thus, interpretations of performance on constructed-response items that are based on the assumption that they are a more valid measure of "true" reading skill, must be carefully weighed within the context of the lower reliability of those items.

3. The concern with assessing student reading performance across different text types appears to be well-founded, with the evidence suggesting that the relative difficulty of text types vary with the reading ability of the student. The differences in the relationship between text type and reading ability between grades 8 and 10 suggest that there may be developmental or curriculum-based factors that bear further scrutiny.

Possible Methodological Issues and Limitations

Data Characteristics

Each of the statistical methods used in this study is based on a set of assumptions regarding characteristics of the data and sample size. Small sample sizes can pose serious problems, particularly for factor analytic methods. Fortunately, such problems were not a concern due to the large number of subjects considered in this study (N = 10,620 for Grade 8, N = 9,068 for Grade 10). In fact, in some cases (such as using fit indices based on the chi-square statistic), I obtained statistically significant results that could be interpreted as primarily reflecting a large number of subjects rather than effect size.

One assumption that is common to all of the methods used is normal distribution of the data. As is true with other studies similar to this one, the data characteristics of test results meant that this assumption could not be met. Multiple-choice items are scored zero (incorrect) or one (correct), and thus do not possess a normal distribution. Constructed-response items scored on a 0-3 or 0-4 scale may approach normality, but also suffer from a limited range of possible scores. Studies of the dimensionality of test data using methods based on correlation matrices have been found to produce varying results, perhaps in part due to the characteristics of the data (DeMars, 2003). For this study, violations of data normality were tolerated because the particularly large sample sizes on which the analyses were based provide a degree of robustness to non-normality for the results (Loehlin, 2004). Nevertheless, data non-normality must be considered a limitation of this study.

One way of dealing with data non-normality involves parceling items together either by summing or averaging them, and then performing the analysis on those parcels. Parceling is widely used in structural equation modeling and confirmatory factor analysis to improve the characteristics of the variables for those analyses (Bandalos, 2002); such improvements include higher reliability, a higher percentage of common variance, and a lower ratio of variables to subjects (or cases) when the sample size is relatively small (Little, Cunningham, Shahar, & Widaman, 2002; Nasser & Takahashi, 2003).

The decision to use parcels, and especially the method selected by a researcher to create parcels, involve several psychometric and philosophical issues that should be explicitly considered in order to assess the implications of parceling for interpreting the results of the analysis. Kim and Hagtvet (2003) argue that "the use of item parcels and their advantages are theoretically justifiable and meaningful only when the parcels retain the information acquired from the items" (p. 102). Recent reviews of the parceling literature indicate a general consensus that parcels should be assembled of items that share one source of common variance to avoid obscuring the true pattern of variation of the original items (Bandalos, 2002; Little et al., 2002). From the general perspective of the design and theoretical focus of a study, Little et al. (2002) propose that parceling not be used in situations where a researcher wants to capture the full complexity of variance at the item level, including loadings on multiple factors and correlated error variance. In contrast, "parceling is more strongly warranted" when the focus is on the latent factors and the relationships among them rather than item-level information (p. 169). In the case of the HSA, since each item is hypothesized to be an indicator of the test taker's ability in the given strand it is important to retain as much item-level information as possible. Hence, a parceling strategy was deemed inappropriate.

Measuring the Reading Process with Large-Scale, Standardized Tests

The limited psychometric evidence for the DOE's three reading strands mirrors the general difficulty that researchers have had in finding evidence for reading components in the results of large-scale tests. One could argue that this lack of evidence indicates that the reading components do not exist, and that the HSA Reading test data accurately reflect this. However, as we have seen there is evidence from various theoretical perspectives and methods that suggests that reading is, in fact, a multi-component skill. Furthermore, the process of aligning test questions with strands involved a close examination of the questions by professionals with expertise and experience in reading instruction and curriculum design. Discounting both of these sources of support of reading as a multidimensional ability should be done carefully, and perhaps only with more evidence than is provided in this study.

An alternative scenario is that reading components do exist, but did not emerge from the data. If we proceed from this position, then several possible methodological issues can be raised in explaining the lack of psychometric evidence emerging from the HSA data. One possibility suggested by previous research is that reading components are difficult to ascertain with a large-scale, standardized instrument like the HSA Reading test. As noted in Chapter 2 of this study, many of the studies that have found evidence for reading components have not been based on large-scale test results but rather on close analyses of the reading process. It has been noted that measuring the *products* of reading (multiple-choice selections, essays) is an imperfect way of assessing the *process* of reading. If this problem is significant, then large-scale tests, by their very nature, can only provide us with limited and indirect information on the reading process. Interestingly, however, it has been pointed out that the movement towards standards-based testing has promoted the development of a more process-oriented understanding of what students should know and be tested on (Linn, 2002a). Ironically, then, as content standards reflect an increasing emphasis on cognitive processes, they may be making it more difficult for large-scale tests to accurately measure those standards.

Proponents of performance-based assessments point to this inability of large-scale, standardized tests to directly measure the process of reading as one of several important limitations of tests, and there is no lack of suggested alternative assessment methods (Murphy, 1998). For example, Fuchs et al. (2001) have argued that oral reading fluency is the most accurate measure of reading ability, while another study (Stage & Jacobsen, 2001) has shown that fluency reliably predicted performance on a standards-based state reading assessment for fourth graders. Other researchers, while acknowledging the demonstrated strong correlation between fluency and other measures of reading ability, caution that the link between the two needs further investigation and that immediate use of fluency in classroom assessment may provide misleading results (Tindal & Marston, 1996). If our concern is with the process of reading, oral reading fluency may represent a useful but limited measure, in that the method used to obtain information occurs simultaneously during the act of reading. Whether oral fluency or other performance-based assessment methods can be adapted to a large-scale, standards-based testing model, however, remains a difficult question (Baker, O'Neil, & Linn, 1993; Linn, Baker, & Dunbar, 1991). Any such adaptation is further complicated by research that has found performance-based assessments to suffer from low reliability, making it difficult to study their construct validity with respect to other measures of reading (Crehan, 2001).

Accounting for the Context of Reading

A related possible problem might be a lack of context variables in the analysis. As discussed in Chapter 2, several researchers argue that context variables are essential in assessing reading, one of the most important of which is the reader's background knowledge (Shapiro, 2004). Simply stated, without knowing how a reader's background knowledge facilitates or hinders comprehension of a particular reading passage, it is difficult to make inferences about reading ability based on test scores. Of related interest in this regard are the numerous possible variables related to the reader's affect—for example, motivation, anxiety, high stakes—that may play important roles in understanding components or skills that may be apparent only under certain circumstances (Roeser et al., 2002).

Further investigation into the role that background or prior knowledge plays in reading performance on passage-based reading tests might focus on the issue of local item dependence (LID) among items tied to a common passage. LID occurs when the content of one item (for example, any of the questions linked to the same reading passage) provides assistance in answering another question. In such cases, the ability to correctly answer a question reflects not only the underlying ability it is intended to measure (i.e., reading), but also information gained from other questions. Indications that significant LID effects exist among items linked to the same passage might be an indication that the performance on those items may be influenced by the content of the passage, rather than only, or even mainly, reading ability. In turn, the strength of the passage effect may be a reflection of the reader's prior knowledge. Interestingly, although LID has generally been seen as a problem in test construction because of its negative effects on score reliability, it has been noted (Lee, 2002) that such effects might be worth the improved construct validity that results from the ability of passages to elicit "a wider range of comprehension sub-skills and process" (p. 12).

Multiple Sources of Measurement Problems

It is possible that these and other factors interact and hinder a psychometric confirmation of reading components. Efforts to assess reading components with a test such as the HSA are likely complicated by the following conditions: (a) any method of assessment of reading ability must contend with item format effects, prior knowledge effects, and various other test circumstance effects; (b) theoretical perspectives on the competencies that compose reading ability range widely, and constitute a "moving base" upon which to build the content standards that, in turn, inform test item construction; and (c) the very response patterns that might be expected from items reflecting different components might be obscured due to a test development process informed by item response theory, which assumes in most circumstances construct unidimensionality among all items. It is understandable, given these conditions, that the task of aligning test items and content standards would be a difficult matter, as demonstrated by the relative lack of psychometric evidence for reading strands. For similar reasons, including discrepancies relating to item format and alignment between standards and items, Linn (1998) has argued that the NAEP results should be used with caution as indicators of achievement.

Finally, it should be noted that the selection of a single-factor model of reading ability as the preferred model reflects its parsimony relative to multiple-factor models with comparable measures of fit to the HSA data. However, it has been argued that a statistically parsimonious model may not be the preferred model when considering other aspects of test structure, such as content representation, student preparation for the test, and development of test items (Meara & Sireci, 1999). Thus, this analysis should not be interpreted as a complete assessment of the strengths and weaknesses of a multiple-component model of reading ability.

Operationalizing Components and Aligning Items

The NAEP's caveat that its reading "aspects" should not be used for diagnostic purposes relieves it of what is perhaps the central burden posed by attempting to identify components of reading: the diagnostic implications of aligning items with components for understanding actual variations in reading ability. But it is fair to ask, then, what the point is of building a reading test upon content standards that specify reading components if those components have no diagnostic function? Or, stated in another way, what does alignment of items and standards accomplish if performance on those items cannot be used as indicators of proficiency at the underlying standard? If reading components are going to be interpretable and acceptable to test takers and those who make decisions based on test results, then it would seem that components must have a role in the diagnosis of reading ability. And if they are going to be interpretable and acceptable within the context of a specific conceptualization of reading ability as defined by those components, then the diagnostic links between test items and reading components needs to be clear and explicit. Otherwise, a single reading score representing overall performance would convey the information necessary for students, parents, and educators.

A possible answer to this problem is suggested by Bhola, Impara, and Buckendahl (2003), who note that there may a mismatch between the specificity of content standards and assessment strategies that better fit more holistic interpretations of the ability being measured. It is widely understood that content standards need to be specific in a number of ways in order to be useful: in particular, they should be specific about the ability and knowledge a student should possess and about the dimensions or levels of that ability and knowledge that are appropriate at different grade levels or clusters (American Educational Research Association, 2003; Rothman et al., 2002; Wixon et al., 2002). This imposition of specificity in delineating ability and knowledge, however, may not be appropriate for all subject areas. In this sense, perhaps the NAEP reading "aspects" reflect theoretical specificity that may not be directly translated into skills that can be assessed or conceptualized as distinct. In fact, as the wording of the NAEP caveat suggests, it may be that the skills to which the "aspects" refer are interpretable and assessable only as parts of a whole. Although the implications of this admittedly nebulous concept of component skills are not entirely clear, they certainly include caution in using a standardized test to discern the presence and level of those skills.

A related problem may lie in the processes used to develop content standards and align items with them. As noted in the discussion in Chapter 2, standards development and alignment are generally performed by curriculum and content experts. The face validity of this process is clear, and there is no reason to believe that such experts should have a lesser role. However, one concern that has been raised regarding the use of experts is their qualitatively different approach to content within their domain of expertise versus that of novices (Ayala, Shavelson, Yin, & Schultz, 2002). Specifically, "experts are consistent in their substantive representations of the principle underlying the [performance] task, whereas novices are strongly influenced by the surface features of the task" (p. 110). Thus, experts might be "expected to process the test items somewhat differently from target test-takers" (Alderson, 2000, p. 97). If this finding can be applied to the processes by which content standards are developed and test items aligned with those standards, it might indicate that the reasoning behind the answers to those items may suffer from more context variance than expected, and thus reflect constructs and abilities unrelated to reading.

Are Content Standards Held to Norm-Referenced Criteria?

Standards-based assessment programs are intended, in part, to establish a more appropriate metric for measuring student achievement than the performance of comparable students across the nation. Standards-based scores are still interpreted relative to something, of course, but that "something" is now a set of statements about knowledge and ability that embody the knowledge and abilities that the student's community has determined to be the goals of education. The very term "standards-based" suggests that assessment is tied to an interpretive framework that is anchored on ground that is more stable and enduring than norms that have been derived from the performance of all test takers. Further, the specificity of the standards provides the means by which test scores can yield more information about the student's knowledge and abilities than simply his or her performance relative to other students. We are able to say that student A is more proficient in reading than student B with respect to aspects X and Y of reading, but is less proficient with respect to aspect Z. The importance of this additional information cannot be denied; with it, educators have a far better base on which to make decisions about curriculum, instruction, and student performance.

It is evident that the promise of standards-based assessment depends crucially on the quality of its standards. The question arises, then: How can we best evaluate the quality of those standards? This study has attempted one such evaluation from a psychometric perspective. The HSA language arts standards were recently subjected to another kind of assessment. According to the State Auditor's report (Hawai'i State Auditor, 2001), the content standards of the HCPS II (as the HSA was then called) were evaluated according to three criteria: comprehensiveness, rigor, and specificity. Accordingly, the questions asked of the content standards were, respectively: (a) "Do the standards address significant concepts and skills for each subject area?" (b) "Are concepts and skills presented at the appropriate level of difficulty?" and (c) "Are the content and skills described specifically enough to be meaningful?" (p. 22). These questions were then addressed by consulting "reference documents" that had been developed by Mid-continent Research for Education and Learning (McREL). As described by the State Auditor, those reference documents were used to evaluate the content standards in the following manner:

> These studies [the reference documents] identify the knowledge and skills that are consistently found within and across highly rated state standards documents and significant national documents in the subject areas. Therefore, these studies were used in the assessment of the comprehensiveness of the Department of Education's standards. The reference documents were also used to determine the appropriate grade level or grade-cluster placement of the benchmarks. If concepts or skills are placed at an earlier grade cluster in the Department of Education's standards than is common within the reference documents, then the department's standards could be said to be more challenging or more rigorous for students. If

> students in Hawai'i are not expected to master content until well after their peers in other states, the department's standards could be said to be less challenging or less rigorous. Finally, the appropriate level of specificity, or detailed description of content, was likewise determined by comparing the content description in the Department of Education's standards against those of the reference documents (Hawai'i State Auditor, 2001, p. 23).

In this case, then, the DOE's content standards appear to have been evaluated by an essentially norm-referenced process. The norms were based on standards documents of five states that were deemed to contain "exemplary content" in the respective subject areas. As described earlier in Chapter 3, the three studies that had been used to select those five states were based on different selection methods and criteria. The extent to which those methods and criteria included psychometric research is difficult to ascertain from the available sources. If such research was used, however, it was not accorded a prominent role in the stated rationale for the selection criteria. It is worth considering the possibility, then, that psychometric research has had a limited role in forming and evaluating reading content standards. If so, more studies such as this one might be useful in promoting the fair use of large-scale test results by holding tests to the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999), thereby continuing our efforts to improve the way we assess student learning.

Conclusion

The widespread use of large-scale, standards-based testing to measure educational achievement and, by extension, instructional effectiveness, administrative efficiency, and progress towards policy goals, has created a clear need for ongoing empirical study to validate the appropriate uses of the resulting test scores. An important part of these validation efforts concerns the relationship between the content standards that inform the structure and content of the test, and the scores generated by the test questions in actual test administrations. The inaugural HSA test data provide us with an opportunity to look carefully at a set of content standards and their relationship to the test scores that constitute an essential part of the context within which those standards should be understood. Analysis of those content standards adds to our knowledge of how standards-based test results should be understood and interpreted.

The results of this study suggest that we may need to be more conservative in our expectations of the kinds of assessment information we can validly obtain from large-scale, standards-based reading tests. This conclusion is at odds with the premise that the content standards that underlie standards-based tests must comprehensively represent the domain of interest in order to be useful as a basis for assessment. In fact, the results of this study indicate that efforts to link large-scale tests to content standards that are comprehensive in terms of the breadth and depth of the reading domain may lead to interpretations of test results that cannot be psychometrically supported.

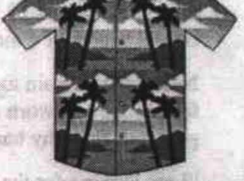
APPENDIX. SAMPLE HSA READING TEST QUESTION

The reading passage and associated test questions on the following pages are taken from the *Teacher's Guide for Interpreting the Hawai'i Content and Performance Standards* (Hawai'i Department of Education, 2003). According to the *Teacher's Guide*, all sample items included in the guide are taken from a "live administration" the HSA and are being publicly released "to provide additional information regarding the kinds of knowledge and skills that students are expected to demonstrate on HCPS II State Assessment" (p. 28).

RECEIVED
AS
FOLLOWS

The Aloha Shirt

By Sophia Schweitzer



1 Not even tucked in, blazing and bold, here's the aloha shirt!

2 The missionaries might have denounced extravagance and nakedness, but the craving for vivid colors, tropical textures, and sensual shapes couldn't be suppressed. Within two centuries a modest workman's shirt grew into the trademark wear of Hawaii.

3 But the history of this shirt reflects the growing pains of a nation and the true Hawaiian aloha shirt has become increasingly hard to find.

4 In the late 1920s and early 1930s tourists, always looking for exotic souvenirs, fell for a fad of the young islanders, unusual prints. Artists and tailors spotted a serious business. The name "aloha shirt," registered in 1936, soon labeled a flourishing industry.

5 Paintings of famous artists were transferred to the fabric of choice, rayon, silkier than silk and inexpensive. Designs competed in intricacy. Border shirts, picture shirts, patterned shirts. How many ways to say Hawaii? Labels themselves became works of art, reflecting inspiration and wild dreams of success.

6 After the darkness of the second World War, colorful, exotic prints were more than ever what visitors wanted. Add to this the attention Hawaii received in the 1950s when it competed with Alaska to become the 49th state, as well as the intrigue with Hollywood. Aloha shirts became a craze.

7 Elvis Presley, John Wayne, Frank Sinatra, all going Hawaiian. Montgomery Clift, dead in a ditch in "From Here to Eternity," in Hawaiian print. Immortal, from now on. Endorsements by world-famous gold-medal swimmer and master surfer Duke Kahanamoku. Photographs of presidents.

8 Kamehameha Garment Company, one of the largest pioneer manufacturers, shipped 35 tons of garments to the mainland in 1960. "Made in Hawaii" sells!

9 In Hawaii, during those crazy years, opinions varied. All good and well in leisure time, but what about business? Many companies fought the breezy aloha shirt.

10 "Spiritually destructive," said a Japanese boss in 1955. "Truth is" writes Honolulu Magazine, in 1967, "almost no man past 30 really looks good in an aloha shirt."

11 Questionable as this might be, with the large demand on the mainland came the need for more effective production. First factories took over. Then labor and designers overseas. On the mainland, imitation shirts appeared. Designs lost their artistic quality. Matching his and hers, no inspiration. Demand lowered, prices dropped. The shirt became "tacky."

12 Only one company in the whole state of Hawaii decided to stay true to the original Hawaiian shirt. Reyn Spooner, created in 1956, is the only aloha shirt line designed and produced right here in Hawaii, its prints still pulsating with local strength.

13 Otherwise weakened, the Hawaiian shirt lost uniqueness. Yet the greatest loss is that no one thought of keeping track. Numerous designs have vanished in the cotton clouds of history. Original shirts with original labels have become collectors items worth hundreds and even thousands of dollars.

14 The true Hawaiian shirt reads like a painting of paradise. The fabric is a canvas for the rich images of the islands. And aloha shirts, true or not, are here to stay. Still the greatest souvenir. They will forever mirror what Hawaii is about. The challenge is now to find the real one, the one that shows aloha.

"The Aloha Shirt" by Sophia Schweitzer, from the Coffee Times Website.

COMPREHENSION PROCESSES STRAND:

Use reading strategies within the reading process to construct meaning.

1 If you wanted to remember what information is in the article, what strategy would be *most* effective?

- A Creating an outline for the article
- B Making a list of nouns found in the article
- C Drawing a picture of an aloha shirt based on the article
- D Listing questions that are not answered in the article

CONVENTIONS AND SKILLS STRAND:

Apply knowledge of the conventions of language and texts to construct meaning

2 What organizational structure does the author use *mostly* in the article?

- F Cause and effect
- G Chronological order
- H Spatial order
- J Question and answer

RESPONSE STRAND:

Respond to texts from a range of stances: initial understanding, personal, interpretive and critical.

3 What does the author cite as proof of the growing popularity of Hawaiian shirts in the 1950s?

- A Their adoption by Hollywood stars
- B Their modest origins
- C Their silky fabrics
- D Their mention in a magazine

CONVENTIONS AND SKILLS STRAND:

Apply knowledge of the conventions of language and texts to construct meaning

4 Read this sentence from paragraph 14 of the article.

> **The true Hawaiian shirt reads like a painting of paradise.**

What literary device does the author use in this sentence to show the artistic beauty of the Hawaiian shirt?

- F Personification
- G Irony
- H Rhyme
- J Simile

Answer Key: 1-A; 2-G; 3-A; 4-J

RESPONSE STRAND:
Respond to texts from a range of stances: initial understanding, personal, interpretive and critical.

5 Which fact supports the idea that "the Hawaiian shirt has lost uniqueness"?

A Hawaiian shirts became a fad in the late 1920s.
B Hollywood stars started wearing Hawaiian shirts.
C Many designs were implemented in the Hawaiian shirts.
D Factories produced Hawaiian shirts in mass quantity.

RESPONSE STRAND: **OPEN-ENDED RESPONSE ITEM:**
Respond to texts from a range of stances: initial understanding, personal, interpretive and critical.

6 What details does the author use to support the idea that aloha shirts became a craze?

RESPONSE STRAND: **EXTENDED-RESPONSE ITEM:**
Respond to texts from a range of stances: initial understanding, personal, interpretive and critical.

26 Write a note to a friend explaining why some collectors are willing to spend hundreds of dollars on aloha shirts.

Answer Key: 5-D; 6 and 26 see student responses later in this guide

REFERENCES

Alderson, J. C. (1990). Testing reading comprehension skills (part one). *Reading in a Foreign Language, 6*(2), 425-438.

Alderson, J. C. (2000). *Assessing reading.* Cambridge, UK: Cambridge University Press.

American Educational Research Association. (2003, Spring). Standards and tests: Keeping them aligned. *Essential Information for Education Policy, 1,* 1-4.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Federation of Teachers. (2001). *Making standards matter 2001: A fifty-state report on efforts to implement a standards-based system.* Retrieved September 1, 2004, from http://www.aft.org/pubs-reports/downloads/teachers/msm2001.pdf

Anderson, R. C. (2004). Role of the reader's schema in comprehension, learning, and memory. In R. B. Ruddell & N. J. Unrau (Eds.), *Theoretical models and processes of reading* (5th ed., pp. 594-606). Newark, DE: International Reading Association.

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic processes in reading comprehension. In P. D. Pearson, R. Barr, M. L. Kamil & P. B. Mosenthal (Eds.), *Handbook of reading research* (pp. 255-291). New York: Longman.

Arizona Department of Education School Effectiveness Division. (2003). *Arizona* *academic content standards, language arts standard 1, reading standard articulated by grade level*. Retrieved September 1, 2004, from http://www.ade.state.az.us/standards/language-arts/articulated.asp

Ayala, C. C., Shavelson, R., Yin, Y., & Schultz, S. (2002). Reasoning dimensions underlying science achievement: The case of performance assessment. *Educational Assessment, 8*(2), 101-121.

Ayala, C. C., Yin, Y., Schultz, S., & Shavelson, R. (2002). *On science achievement from the perspective of different types of tests: A multidimensional approach to achievement validation. CSE technical report.* Los Angeles, CA: Center for the Study of Evaluation; Center for Research on Evaluation, Standards, and Student Testing.

Baker, E. L., O'Neil, H. F., Jr., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist, 48*(12), 1210-1218.

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling, 9*(1), 78-102.

Beck, I. L., & McKeown, M. G. (2002). Comprehension: The sine qua non of reading. In S. Patton & M. Holmes (Eds.), *The keys to literacy* (pp. 43-51). Washington, DC: Council for Basic Education.

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1-27).

Hillsdale, NJ: Lawrence Erlbaum Associates.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice, 22*(3), 21-29.

Brennan, R. L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice, 17*(1), 5-9.

Bridgeman, B., & Rock, D. A. (1993). Relationships among multiple-choice and open-ended analytical questions. *Journal of Educational Measurement, 30*(4), 313-329.

Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning, 47*(3), 423-466.

Byrne, B. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cain, K., Bryant, P., & Oakhill, J. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology, 96*(1), 31-42.

California Department of Education. (1998). *English-Language Arts content standards for California public schools kindergarten through grade twelve*. Retrieved September 1, 2004, from http://www.cde.ca.gov/re/pn/fd/documents/english-language-arts.pdf

Carlo, M. S., & Sylvester, E. S. (1996). *Adult second-language reading research: How may it inform assessment and instruction? (NCAL Technical Report TR96-08).* Philadelphia, PA: National Center on Adult Literacy, University of Pennsylvania. (ERIC Document Reproduction Service No. ED412373)

Carr, T. H., Brown, T. L., Vavrus, L. G., & Evans, M. A. (1990). Cognitive skill maps and cognitive skill profiles: Componential analysis of individual differences in children's reading efficiency. In B. A. Levy (Ed.), *Reading and Its Development: Component Skills Approaches* (pp. 1-55). San Diego, CA: Academic Press, Inc., Harcourt Brace Jovanovich, Publishers.

Carver, R. P. (1993). Merging the simple view of reading with rauding theory. *Journal of Reading Behavior, 4*, 439-454.

Catts, H. W., Hogan, T. P., & Fey, M. E. (2003). Subgrouping poor readers on the basis of individual differences in reading-related abilities. *Journal of Learning Disabilities, 36*, 151-164.

Chard, D. J., Simmons, D. C., & Kameenui, E. J. (1995). *Understanding the primary role of word recognition in the reading process: synthesis of research on beginning reading (Technical Report No. 15)*. Eugene, OR: National Center to Improve the Tools of Educators, College of Education, University of Oregon. (ERIC Document Reproduction Service No. ED386862)

Crehan, K. D. (2001). An investigation of the validity of scores on locally developed performance measures in a school assessment program. *Educational &*

Psychological Measurement, 61(5), 841-848.

Cunningham, A. E., Stanovich, K. E., & Wilson, M. R. (1990). Cognitive variation in adult college students differing in reading ability. In T. H. Carr & B. A. Levy (Eds.), *Reading and Its Development: Component Skills Approaches* (pp. 129-159). San Diego, CA: Academic Press, Inc., Harcourt Brace Jovanovich, Publishers.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*(1), 55-77.

DeMars, C. E. (2003). Detecting multidimensionality due to curricular differences. *Journal of Educational Measurement, 40*(1), 29-51.

Diegmueller, K. (1994). Panel unveils standards for history: Release comes amid outcries of imbalance. *Education Week, 14*(9), 1, 10.

Duarte Silva, A. P., & Stam, A. (1995). Discriminant analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 277-318). Washington, DC: American Psychological Association.

Easton, J. Q., Correa, M., Luppescu, S., Park, H.-S., Ponisciak, S., Rosenkranz, T., et al. (2003). *How do they compare? ITBS and ISAT reading and mathematics in the Chicago Public Schools, 1999 to 2002.* Chicago, IL: Consortium on Chicago School Research.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed response item types. *Journal of Educational Measurement, 35*(2), 137-154.

Fenton, R. (2003, April). *How have state level standards-based tests related to norm-referenced tests in Alaska?* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Florida Department of Education. (2001). *FCAT test item and performance task specifications*. Retrieved January 28, 2004, from http://www.firn.edu/doe/sas/fcat/fcatis01.htm

Freedle, R., & Kostin, I. (1994). Can multiple-choice reading tests be construct-valid? A reply to Katz, Lautenschlager, Blackburn, and Harris. *Psychological Science, 5*(2), 107-110.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239-256.

Geisinger, K. F. (1992). The metamorphosis of test validation. *Educational Psychologist, 27*(2), 197-222.

Good, R. (2002, April). *Using discriminant analysis as a method of combining multiple measures of student performance.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Gough, P. B., Hoover, W. A., & Peterson, C. L. (1996). Some observations on a simple view of reading. In C. Cornoldi & J. Oakhill (Eds.), *Reading Comprehension Difficulties: Processes and Intervention.* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Grabe, W., & Stoller, F. L. (2002). *Teaching and researching reading.* London: Pearson

Education Longman.

Guthrie, J. T., Wigfield, A., Metsala, J. L., & Cox, K. E. (2004). Motivational and cognitive predictors of text comprehension and reading amount. In N. J. Unrau (Ed.), *Theoretical models and processes of reading* (5th ed., pp. 929-953). Newark, DE: International Reading Association.

Hannon, B., & Daneman, M. (2001). A new tool for measuring and understanding individual differences in the component processes of reading comprehension. *Journal of Educational Psychology, 93*(1), 102-128.

Hatcher, L. (1994). *A step-by-step approach to using the SAS System for factor analysis and structural equation modeling.* Cary, NC: SAS Institute, Inc.

Hawai'i Department of Education. (2001). *Teacher's guide for interpreting the Hawai'i Content and Performance Standards* (2nd, first printed ed.). Honolulu: Hawai'i Department of Education.

Hawai'i Department of Education. (2003). *Teacher's guide for interpreting the Hawai'i Content and Performance Standards* (2nd, second printed ed.). Honolulu: Hawai'i Department of Education.

Hawai'i Department of Education Office of Accountability and School Instructional Support/School Renewal Group. (1999). *Making sense of standards: Moving from the Blue Book to HCPS II* (2nd ed.). Honolulu, HI: Hawai'i Department of Education.

Hawai'i Department of Education Office of Curriculum Instruction and Student Support / Instructional Services Branch. (2003). *Curriculum framework for language arts* *(draft)*. Honolulu: Hawai‘i Department of Education.

Hawai‘i State Auditor. (2001). *A review and assessment of the Department of Education's development of educational standards*. Honolulu, HI: Hawai‘i State Office of the Auditor.

Hawai‘i State Performance Standards Review Commission. (2003). *Final report*. Honolulu, HI: Hawai‘i State Department of Education.

Hawaii State Department of Education. (2001). *Teacher's guide for interpreting the Hawaii Content and Performance Standards* (Second ed.). Honolulu: Hawaii Department of Education.

Johnston, P. H. (1984). Assessment in reading. In P. B. Mosenthal (Ed.), *Handbook of reading research* (pp. 147-182). New York: Longman.

Katz, S., & Lautenschlager, G. J. (2001). The contribution of passage and no-passage factors to item performance on the SAT Reading task. *Educational Assessment*, *7*(2), 165-176.

Kendall, J. S., Snyder, C., Schintgen, M., Wahlquist, A., & Marzano, R. J. (1999). *A distillation of subject-matter content for the subject-areas of language arts, mathematics, and science*. Aurora, CO: Mid-continent Research for Education and Learning.

Kifer, E. (2001). *Large-scale assessment: Dimensions, dilemmas, and policy*. Thousand Oaks, CA: Corwin Press.

Kim, S., & Hagtvet, K. A. (2003). The impact of misspecified item parceling on representing latent variables in covariance structure modeling: A simulation study. *Structural Equation Modeling, 10*(1), 101-127.

LaBerge, D., & Samuels, S. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*, 293-323.

Lee, Y. (2002). *Score reliability of a test composed of passage-based testlets: A generalizability theory perspective.* Paper presented at the First International Conference of the Korea English Education Society, Chungbuk, Korea.

Lennon, R. T. (1962). What can be measured? *Reading Teacher, 15*, 326-337.

Levy, B. A., & Carr, T. H. (1990). Component process analyses: Conclusions and challenges. In T. H. Carr & B. A. Levy (Eds.), *Reading and its development: Component skills approaches* (pp. 423-438). San Diego, CA: Academic Press, Inc., Harcourt Brace Jovanovich.

Li, Y. H. (2001, April). *An evaluation of the construct validity for the multiple-subject testing programs.* Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Li, Y. H., Ford, V., & Tompkins, L. J. (1999, April). *The construct validity of a performance-based assessment program.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Linn, R. L. (1998). Validating inferences from National Assessment of Educational Progress achievement-level reporting. *Applied Measurement in Education, 11*(1), 23-47.

Linn, R. L. (2002a). Constructs and values in standards-based assessment. In H. I. Braun, D. N. Jackson, D. E. Wiley & S. Messick (Eds.), *The role of constructs in* *psychological and educational measurement* (pp. 231-254). Mahwah, NJ: Lawrence Erlbaum Associates.

Linn, R. L. (2002b). Validation of the uses and interpretations of results of state assessment and accountability systems. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 27-48). Mahwah, NJ: Lawrence Erlbaum Associates.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Linn, R. L., & Hambleton, R. K. (1991). Customized tests and customized norms. *Applied Measurement in Education, 4*(3), 185-207.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*(2), 151-173.

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Lubliner, S., & Smetana, L. (2003, April). *Recognition or recall: What reading comprehension tests really measure.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Lyon, G. R. (2002). Overview of reading and literacy research. In M. Holmes (Ed.), *The keys to literacy* (pp. 8-16). Washington, DC: Council for Basic Education.

MacQuarrie, D. (2003, April). *Validity evidence for Washington Assessment of Student Learning (WASL) performance standard cut-scores for reading and mathematics.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207-218.

Maruyama, G. M. (1998). *Basics of structural equation modeling.* Thousand Oaks, CA: SAGE Publications, Inc.

Massachusetts Department of Education. (2001). *Massachusetts English language arts curriculum framework.* Retrieved January 28, 2004, from http://www.doe.mass.edu/frameworks/ela/0601.pdf

McCabe, P. P. (2003). Enhancing self-efficacy for high-stakes reading tests. *Reading Teacher, 57*(1), 12-20.

McCandliss, B., Beck, I. L., Sandak, R., & Perfetti, C. A. (2003). Focusing attention on decoding for children with poor reading skills: Design and preliminary tests of the word building intervention. *Scientific Studies of Reading, 7*(1), 75-104.

Meara, K., & Sireci, S. G. (1999). *Appraising the dimensionality of the Medical College Admissions Test*: Association of American Medical Colleges.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 61-75). Hillsdale, NJ: Lawrence Erlbaum Associates.

Murphy, S. (1998). *Fragile evidence: A critique of reading assessment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Nasser, F., & Takahashi, T. (2003). The effect of using item parcels on ad hoc goodness-of-fit indexes in confirmatory factor analysis: An example using Sarason's Reactions to Tests. *Applied Measurement in Education, 16*(1), 75-97.

National Assessment Governing Board. (2002). *Reading framework for the 2002 National Assessment of Academic Progress.* Retrieved April 21, 2004, from http://www.nagb.org/pubs/reading_Framework/ch4.html#1

National Assessment Governing Board. (2003). *Reading framework for the 2003 National Assessment of Educational Progress.* Retrieved April 20, 2004, from http://www.hagb.org/pubs/reading_Framework/ch2.html

O'Neil, H. F., Jr., & Brown, R. S. (1998). Differential effects of question formats in math assessment on metacognition and affect. *Applied Measurement in Education, 11*(4), 331-351.

Patton, S., & Holmes, M. (Eds.). (2002). *The keys to literacy.* Washington, DC: Council for Basic Education.

Pearson, P. D., Barr, R., Kamil, M. L., & Mosenthal, P. B. (1984). *Handbook of reading research.* New York: Longman.

Pearson, P. D., & Garavaglia, D. R. (2003). *Improving the information value of* *performance items in large scale assessments.* Washington, DC: National Center for Education Statistics.

Pellegrino, J. W., Chudowsky, N., Glaser, R., & National Research Council (U.S.). Division of Behavioral and Social Sciences and Education. Committee on the Foundations of Assessment. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Perfetti, C. A. (1988). Verbal efficiency in reading ability. In M. Daneman, G. E. Mackinnon & T. G. Waller (Eds.), *Reading research: Advances in theory and practice* (Vol. 6, pp. 109-143). New York: Academic Press.

Perfetti, C. A., Marron, M. A., & Foltz, P. W. (1996). Sources of comprehension failure: Theoretical perspectives and case studies. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and remediation* (pp. 137-165). Mahwah, NJ: Lawrence Erlbaum Associates.

Perkins, K., & Pohlmann, J. T. (2002). *Changes in the responses to an English as a Second Language reading comprehension test (Report No. TM034628).* (ERIC Document Reproduction Service No. ED 471462)

Potts, G. R., & Peterson, S. B. (1985). Incorporation versus compartmentalization in memory for discourse. *Journal of Memory and Language, 24*, 107-118.

Pullen, P. C., & Justice, L. M. (2003). Enhancing phonological awareness, print awareness, and oral language skills in preschool children. *Intervention in School and Clinic, 39*(2), 87-98.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40*(2), 163-184.

Roeser, R. W., Shavelson, R. J., Kupermintz, H., Lau, S., Ayala, C., Haydel, A., et al. (2002). The concept of aptitude and multidimensional validity revisited. *Educational Assessment, 8*(2), 191-205.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing.* Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing.

Rupley, W. H., Willson, V. L., & Nichols, W. D. (1998). Exploration of the developmental components contributing to elementary school children's reading comprehension. *Scientific Studies of Reading, 2*(2), 143-158.

Samuels, S. J., & Kamil, M. L. (1984). Models of the reading process. In P. D. Pearson, R. Barr, M. L. Kamil & P. B. Mosenthal (Eds.), *Handbook of reading research* (pp. 185-224). New York: Longman.

SAS Institute Inc. (1999). *The STEPDISC procedure, overview.* Retrieved September 11, 2004, from http://www.id.unizh.ch/software/unix/statmath/sas/sasdoc/stat/chap60/sect1.htm

SAS Institute Inc. (1999-2001). The SAS system for Windows (Version Release 8.2). Cary, NC: SAS Institute Inc.

Schedl, M., Gordon, A., Carey, P. A., & Tang, K. L. (1996). *An analysis of the dimensionality of TOEFL reading comprehension items (TOEFL Research*

Reports No. 53). Princeton, N.J.: Educational Testing Service. (ERIC Document Reproduction Service No. ED400327)

Schwartz, S. (1984). *Measuring reading competence: A theoretical-prescriptive approach*. New York: Plenum Press.

Shapiro, A. M. (2004). How including prior knowledge as a subject variable may change outcomes of learning research. *American Educational Research Journal, 41*(1), 159-189.

Singer, H., Ruddell, R. B., & Ruddell, M. R. (Eds.). (1994). *Theoretical models and processes of reading* (4th ed.). Newark, DE: International Reading Association.

Smith, F. (1994). *Understanding reading: A psycholinguistic analysis of reading and learning to read* (5th ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review, 30*(3), 407-419.

Stanovich, K. E. (1986/2004). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. In N. J. Unrau (Ed.), *Theoretical models and processes of reading* (5th ed., pp. 454-516). Newark, DE: International Reading Association.

Stanovich, K. E. (1994). Constructivism in reading education. *Journal of Special Education, 28*(3), 259-274.

Stanovich, K. E. (2000). Concepts in developmental theories of reading skill: Cognitive resources, automaticity, and modularity. In *Progress in understanding reading:*

Scientific foundations and new frontiers (pp. 221-241). New York: The Guilford Press.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stotsky, S. (1997). *State English standards.* Washington, DC: Thomas B. Fordham Foundation.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston, MA: Allyn and Bacon.

Texas Education Agency. (2003). *Interpreting assessment reports: Texas Student Assessment Program.* Retrieved January 28, 2004, from http://www.tea.state.tx.us/student.assessment/resources/guides/interpretive/general.pdf

Thompson, B. (2000). Ten commandments of structural equation modeling. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 261-283). Washington, DC: American Psychological Association.

Tindal, G. (2002). Large-scale assessments for all students: Issues and options. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 1-24). Mahwah, NJ: Lawrence Erlbaum Associates.

Tindal, G., & Marston, D. (1996). Reflections on "Technical adequacy of alternative reading measures as performance assessments". *Exceptionality, 6*(4), 247-251.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 29-44). Hillsdale, NJ: Lawrence Erlbaum Associates.

Virginia Department of Education. (2003). *English Standards of Learning curriculum framework.* Retrieved September 1, 2004, from http://www.pen.k12.va.us/VDOE/Instruction/English/englishCF.html

Wang, X. B. (1999). *Understanding psychological processes that underlie test takers' choices of constructed response items.* Princeton, NJ: Law School Admission Council. (ERIC Document Reproduction Service No. ED467813)

Washington Office of Superintendent of Public Instruction. (2000). *Test specifications for the Washington Assessment of Student Learning.* Retrieved January 28, 2004, from http://www.k12.wa.us/assessment/WASL/reading10rdspecs.aspx

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education (Research monograph No. 6).* Washington, DC: Council of Chief State School Officers. (ERIC Document Reproduction Service No. ED414305)

Whitehurst, G. J., & Lonigan, C. J. (1998). Child development and emergent literacy. *Child Development, 69*, 848-872.

Wisconsin Department of Public Instruction. (1998). *Wisconsin model academic standards.* Retrieved September 13, 2004, from http://www.dpi.state.wi.us/standards/elaa8.html

Wixon, K. K., Fisk, M. C., Dutro, E., & McDaniel, J. (2002). *The alignment of state standards and assessments in elementary reading. CIERA Report No. CS511905.* Ann Arbor, MI: Center for the Improvement of Early Reading Achievement. (ERIC Document Reproduction Service No. ED474625)