PERFORMANCE ON LARGE-SCALE SCIENCE TESTS:

ITEM ATTRIBUTES THAT MAY IMPACT

ACHIEVEMENT SCORES

by

Janet Victoria Gordon

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctorate of Education

in

Education

MONTANA STATE UNIVERSITY
Bozeman, Montana

April 2008

APPROVAL

of a dissertation submitted by

Janet Victoria Gordon

This dissertation has been read by each member of the doctorate committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Division of Graduate Education.

Dr. Jayne Downey

Approved for the Department of Education

Dr. Robert Carson

Approved for the Division of Graduate Education

Dr. Carl A. Fox

STATEMENT OF PERMISSION TO USE

Janet Victoria Gordon

April 2008

ACKNOWLEDGMENTS

I would like to acknowledge many people for helping me during my doctoral work. My first debt of gratitude goes to my committee chair, Dr. Jayne Downey, for her generous time and commitment. Dr. Downey's academic guidance and encouragement helped me to proceed and complete my dissertation. I am also very grateful for having an exceptional doctoral committee and wish to thank Dr. Elisabeth Swanson, Dr. Art Bangert, and Dr. Jennie Luebeck for their continual support and encouragement. Dr. Swanson's expertise and insightful comments helped guide my research. Dr. Bangert's long discussions on MANOVA and effect size helped me to draw conclusions from the statistical analysis. Dr. Luebeck's dedication to read the dissertation helped provide clarity for the reader in the final document.

I am extremely grateful for the assistance and advice I received from Roy Beven and Cinda Parton at OSPI. They provide unprecedented leadership in science assessment for Washington State. I owe a special note of gratitude to Dr. Mike Trevisan for guiding me toward seminal works within the field of educational measurement and for always being available for thought-provoking discussions.

Finally, I extend many thanks to my friends and family for their encouragement. I am thankful for my husband, Phil, who helped edit my drafts. My son, Sterling, especially had a way of helping me keep my life in proper perspective and balance. I am very grateful to the funding provided by the Center for Learning and Teaching in the West (Principal Investigator, Dr. Elisabeth Swanson). I also extend thanks to Micki McGregor for doing an excellent job with managing the award distribution.

TABLE OF CONTENTS

TABLE OF CONTENTS (CONTINUED)

TABLE OF CONTENTS (CONTINUED)

LIST OF TABLES

Table Page

LIST OF TABLES (CONTINUED)

LIST OF FIGURES

Figure Page

ABSTRACT

Significant differences in achievement among ethnic groups persist on the eighth-grade science Washington Assessment of Student Learning (WASL). The WASL measures academic performance in science using both scenario and stand-alone question types.

Previous research suggests that presenting target items connected to an authentic context, like scenario question types, can increase science achievement scores especially in underrepresented groups and thus help to close the achievement gap. The purpose of this study was to identify significant differences in performance between gender and ethnic subgroups by question type on the 2005 eighth-grade science WASL. MANOVA and ANOVA were used to examine relationships between gender and ethnic subgroups as independent variables with achievement scores on scenario and stand-alone question types as dependent variables. MANOVA revealed no significant effects for gender, suggesting that the 2005 eighth-grade science WASL was gender neutral. However, there were significant effects for ethnicity. ANOVA revealed significant effects for ethnicity and ethnicity by gender interaction in both question types. Effect sizes were negligible for the ethnicity by gender interaction. Large effect sizes between ethnicities on scenario question types became moderate to small effect sizes on stand-alone question types. This indicates the score advantage the higher performing subgroups had over the lower performing subgroups was not as large on stand-alone question types compared to scenario question types. A further comparison examined performance on multiple-choice items only within both question types. Similar achievement patterns between ethnicities emerged; however, achievement patterns between genders changed in boys' favor. Scenario question types appeared to register differences between ethnic groups to a greater degree than stand-alone question types. These differences may be attributable to individual differences in cognition, characteristics of test items themselves and/or opportunities to learn. Suggestions for future research are made.

CHAPTER 1

INTRODUCTION

"Sometimes we prayed for easier tests to make us smarter."
Wright & Stone, p. xi, 1979

Over the last 15 years, considerable discussion about gender and cultural fairness of large-scale assessments has ensued. This discussion has led to a reconceptualization of the design of assessments. Washington State, eager to provide fairness for all students, decided to abandon the commercially available, multiple-choice tests originally used in the state's large-scale science testing program. Instead, a cross-cultural team collaboratively develops the standardized science exam incorporating alternative question types found to favor underrepresented groups. This study will evaluate the effectiveness of these alternative question types and provide insight to further improvements to large-scale standardized science exams.

This chapter will introduce and provide background to the problem, identify the problem statement and purpose, list major questions of the study, and discuss the professional significance of the study. This chapter will also define key terms.

The Need for the Study

The Problem

Anastasi and Urbina (1997) advocated that standardized achievement tests, if well-designed, "can facilitate learning, provide a means of adapting instruction to individual needs, aid in the evaluation and improvement of teaching, provide information on the adequacy with which essential skills and content are taught, and aid in the objective, uniform, and efficient assignment of grades" (p. 477). Conversely, standardized achievement tests can reinforce cultural stereotypes and lower student motivation (Steele & Aronson, 1995), lead to test inaccuracies and student misclassification (Solano-Flores & Nelson-Barber, 2001; McGill-Franzen & Allington, 1993) and create abundant measurement misunderstandings between administrators, teachers, parents, and students (Frisbie, 2005). Thus, the creation of fair and equitable tests for all students, requires an understanding of how key factors such as ethnicity, item response format and question type (scenario versus stand-alone) influence student test performance.

During the past five years, these issues have assumed more significance because of federal legislation passed in 2002 under the No Child Left Behind Act (NCLB). This act requires all public schools to have 100% of students at grade level in reading, writing, and mathematics by Spring 2014. To comply with federal legislation, each state in the United States annually administers a standardized achievement test designed to measure students' knowledge. This federal policy requires reporting of grade level performance by student subgroup or ethnicity. Each subgroup must make annual yearly progress (AYP) sufficient to bring 100% of the students to grade level by the spring of 2014.

Previous research has shown that student ethnicity (Klein, Javanovic, Stecher, McCaffery, Shavelson, Haertel, Solano-Flores & Comfort, 1997), student gender (Zenisky, Hambleton & Robin, 2004), item response format (Shavelson, Baxter, & Pine, 1992; Stecher, Klein, Solano-Flores, McCaffrey, Robyn, Shavelson, & Haertel, 2000; Lawrenz, Huffman & Welch, 2000), task stimulus characteristics (Bennett & Ward, 1993; Snow & Lohman, 1989), and the cultural validity of the assessment tool itself (Solano-Flores & Nelson-Barber, 2000; Swisher & Deyhle, 1987) can influence test performance. Research also suggests some question types and item response formats are more conducive to higher academic achievement in underrepresented populations. For example, students of underrepresented groups achieve higher scores on real world, contextual items arranged in a performance assessment compared to traditional, stand-alone items that are often multiple-choice (Stecher, Klein, Solano-Flores, McCaffrey, Robyn, Shavelson, & Haertel, 2000). There is some evidence that indicates girls tend to achieve higher scores on open-ended questions (Zenisky, Hambleton, & Robin, 2004) and authentic performance assessments (Klein et al., 1997) compared to multiple-choice assessments.

It has been suggested that performance assessments may also reduce achievement gaps between whites and non-whites by providing students with opportunities to show evidence of their knowledge rather than recall facts (Solano-Flores & Shavelson, 1997). Performance assessments can make students' content knowledge and cognitive processing more transparent than traditional item response formats, acknowledging individual variation on the path to competency.

Recent developments in performance assessments are part of a movement to address heightened concerns about fairness, "real world" problem solving, and formative feedback to improve student learning. However, the traditional "hands-on" component is difficult to include in large-scale standardized tests due to lengthy scoring procedures and low inter-rater reliability (Gao, Shavelson, & Baxter, 1994). Today the term "performance assessment" broadly refers to item response formats other than traditional multiple-choice, such as constructed-response, observation, portfolios (Sugrue, 1997) and interactive computer-based assessments (Williamson, Bauer, Steinberg, Mislevy, Behrens, & DeMark, 2004). This operational definition will be used in this investigation.

Over the last several years, the National Assessment of Educational Progress (NAEP) assessments have simulated traditional performance assessments by presenting items in thematic blocks or "real world" scenarios. These scenario question types are composed of items with an extended constructed response format designed to resemble entries in a student's science journal. When NAEP began studying question types and item response formats, mathematics and science NAEP tests included both items grouped into real world scenarios and items not grouped into a scenario, simply called "stand-alone" items hence forth in this paper. NAEP found lower off-task and omission rates in the scenario question types compared to stand-alone question types (Jones, 1992; Silver, Cengiz, & Stylianou, 2000). They hypothesized that the authentic real-world context of the scenario question types increased student motivation and engagement compared to stand-alone question types.

Large-Scale Science Assessment in Washington State

Four years have elapsed since Washington State administered their first standardized science assessment. In the development stage of the first science assessment, science teachers and practitioners reviewed various proposed question types that could be used to pattern the science WASL. One of the proposed question types was the scenario, which organized items into thematic blocks patterned after NAEP science tests. Teachers and practitioners preferred the scenario to other question types for a number of reasons. First, the well-contextualized questions helped ensure students could interpret the problem sufficiently. Second, the questions engaged students in a series of tasks that required a synthesis of concepts to solve the problem. In addition, other question types did not seem to measure students' cognitive abilities for interpreting and solving problems within the science domain that students may encounter in real life. Dr. Cathy Taylor, professor at the University of Washington, assisted in the construction of the first science WASL. She stated, "The goal is for students to build knowledge and skills that are transferable to the real world" (C. Taylor, personal communication, June 2, 2005).

Review and revision of the science WASL is an ongoing process. It begins each year when the Science Assessment Leadership Team (SALT) convenes to construct new items. SALT is a cross-cultural team comprised of science teachers, scientists, science assessment professionals from OSPI and educational testing specialists. SALT begins the session by identifying constructs based on the Washington State's science Grade-Level Expectations (GLEs) for the creation of new test items. This test development by local educators and practitioners is expensive when compared to simply administering commercially available traditional multiple-choice tests. Roy Beven, Director of Science Assessment at Washington State's OSPI, strives for an assessment system that is valid and reliable (R. Beven, personal communication, July 18, 2005.) To be valid, the system must measure student understanding of the science Grade-Level Expectations (GLEs), not simply facts and definitions. Each year, the Science Assessment Group at OSPI and SALT strive to identify constructs representing knowledge that is important enough for all students to know.

In the past, these test items would have been formed into approximately 45 stand-alone items with predominantly a multiple-choice response format that forced the student to respond with discrete bits of knowledge. However, the SALT takes a very different approach and "bundles" the items into scenario-type questions that present problems and issues the same way as in real life. Based on the concept of situated cognition, scenarios provide a context or situation by beginning with a "friendly" story that sets the stage for the items that follow. Scenarios are comprised of five to six items, mostly short and extended constructed response format, with one or two items in multiple-choice response format. Scenarios use graphic organizers for the visual display of information thought to assist students in the process of abstracting important information for the assigned task (Ausebel, 1960). In addition, scaffolding techniques assist the student as the student moves from question to question within the scenario.

Decades of research have studied how culture and ethnicity influence the way people interpret and solve test problems. This research has led to theories of cultural variation in cognition. It is this variation that Solano-Flores (2002) used to develop the notion of "cultural validity" as a measure of test validity across cultures. Solano-Flores and Nelson-Barber (2001) suggested that cultural validity in assessment could be attained by incorporating a socio-cultural perspective in the assessment development phase.

The non-traditional approach to test development and the scenario question type on the science WASL may help create a culturally valid assessment tool. Based on current Differential Item Functioning (DIF) studies conducted by OSPI, the science WASL is working in favor of Native Americans, Hispanics, and African Americans (C. Taylor, personal communication, June 6, 2005.) The departure from the traditional multiple-choice item response formats may allow underrepresented groups greater flexibility in providing evidence of learning.

In summary, research suggests item response format and question type (scenario or stand-alone) contributes to achievement score variance with some formats and question types being more conducive to underrepresented populations. Washington State takes a unique approach to the science WASL development that may improve fairness and cultural validity of the test. First, a cross-cultural team of science practitioners, education professionals, and OSPI assessment specialists convene to write the test rather than using a commercially available test. Second, scenarios, rather than stand-alone question types, provide an engaging, authentic context

"The Assessment Era," as Broadfoot & Black (2004) termed the present, and the decisions about what data to collect, how to collect it, and how to interpret it, has far-reaching impacts on graduation and dropout rates of underrepresented populations (McGill-Franzen & Allington, 1993; Messick, 1989). In 2010, Washington State high school graduates will need passing grades on the 10th grade reading, writing, mathematics, and science Washington Assessment of Student Learning (WASL) tests. This research will help inform the development of fair and equitable assessment for all students by providing increased understanding of how question type influences student achievement on high-stakes tests.

The Achievement Gap

Examination of achievement levels among ethnic groups on the science NAEP and science WASL reveal a common pattern. This pattern, called “the achievement gap,” reveals a large disparity between the percentage of White and non-White students who have achieved grade level performance. Although there has been much emphasis on decreasing this achievement gap, the gap between White and Black, Hispanic or American Indian student scores on the 2005 NAEP assessment did not significantly differ from the 1996 or 2000 NAEP science assessments (Grigg, Lauko, & Brockway, 2006).

The 2005 science NAEP used four levels to classify achievement: below basic, basic, proficient, and advanced. Students at a below basic or basic level have not achieved grade level performance and have not met the standards. Students at a basic level have achieved a partial mastery of the fundamental knowledge and skills needed for grade level performance. Students at a proficient level achieved “solid mastery” and students at an advanced level achieved “superior mastery” of the fundamental knowledge and skills needed for grade level performance.

On the 2005 science NAEP, Whites performed better than all other ethnicities at each grade level (4, 8, 12). At grade 8, for example, 40% of White students were proficient or advanced, while less than 12% of Black, Hispanic, and American Indian students were proficient or advanced (Grigg, Lauko, & Brockway, 2006). The percentage of Black, Hispanic, and American Indian students that did not meet grade level standards were 93%, 89%, and 89%, respectively, while the percentage of White students was 60%.

This pattern is similar to the pattern seen on the 2005 8th-Grade science WASL results. The 2005 science WASL also used four levels to classify achievement: below basic, basic, proficient, and advanced. Forty-two percent of White students were proficient or advanced, while less than 20% of Black, Hispanic and American Indian students were proficient or advanced. The percentage of Black, Hispanic, and American Indian students that did not meet grade level standards were 80%, 79%, and 72%, respectively, while the percentage of White students was 55%.

The Purpose of this Study

The purpose of this study is to test the hypothesis that students' performance on the 2005 8th-grade science WASL does not vary by ethnic or gender subgroup across question type (scenario and stand-alone). First, the study will compare student achievement scores across seven ethnic subgroups (American Indian/Alaska Native, Asian, Black African American, Hispanic, White, Hawaiian Pacific Islander, Multiracial) and two genders (Female, Male). Second, the study will compare achievement scores as assessed by two question types (scenario and stand-alone) to determine if one question type is better suited for underrepresented groups. This study addresses three research questions:

Q1. How do ethnic and gender groups perform on the 8th-grade science WASL?

Q2. How do ethnic and gender groups perform on scenario question types?

Q3. How do ethnic and gender groups perform on stand-alone question types?

Operational Definitions

Construct. Anastasi & Urbina (1997) defined a construct as a theoretical entity developed to organize observable responses into a particular trait, for example reading fluency. Examples of constructs assessed by the science WASL are the concepts and principles of physical, earth, space and living systems and links from principles to conditions of application.

Extended Constructed Item Response Format**.** On the science WASL, items with an extended constructed response format require more than two steps and longer (< 3 sentences) responses than short constructed response items. These items may ask students to develop a viable solution, demonstrate an understanding or process, communicate ideas or results, or show reasoning using complex responses. They may ask for a graph, figure, diagram, and/or table with labels, words, sentences, or equations to support students' reasoning. These questions are worth four points each and partial credit may be awarded.

Item Response Format. In the assessment field, test items are categorized according to the format of the student response they require such as multiple-choice, short or extended constructed response (Sugrue, 1997).

Performance Assessment. A term broadly used term to refer to assessments that do not primarily use a traditional, stand-alone question type with a multiple-choice item response format. Today the term performance assessment can refer to assessments that use authentic, scenario question types that typically include items with an extended constructed response format, as well as observation, portfolio and hands-on investigations.

Principle. When two concepts are linked, a principle is formed. An example of a principle is, "Darker colors transfer more heat energy than light colors; therefore, to keep my dog warmer, I will paint the doghouse brown rather than white."

Question Type. Question type, not to be confused with item response format, refers to either a scenario question type or a stand-alone question type. A scenario question type is similar to the NAEP thematic blocks where several test items are presented in an authentic, real-world context. Most of the items in a scenario question type are either short or extended constructed response format. In contrast, a stand-alone question type is simply one item, usually multiple-choice response format.

Scenario Question Type. Based on the concept of situated cognition, scenario questions provide embedded context by beginning with a story that sets the stage for the proceeding questions. Scenarios are comprised of five to six items that are mainly short or extended constructed response format, but can include multiple-choice response formats. Scenarios vary in number of points possible, depending on the item response format of each item.

Short Constructed Response Format. These items require one or two steps to answer completely. They may ask the student to develop a viable solution, demonstrate an understanding or process, communicate an idea or result, or show reasoning using a figure, diagram, equation, and/or a few explanatory sentences. These items include Enhanced Multiple-Choice (EMC) items that ask students to select from a list of four possible responses and then justify or explain the reason(s) for choosing that response. These questions are worth two points each and partial credit may be awarded.

Stand-Alone Question Type. A single item, not grouped by theme or within a scenario question. These items are either a short-answer or multiple choice response format and are worth 2 points each. No partial credit is awarded for stand-alone question types.

Significance of this Study

There are many reasons why it is important to investigate how student achievement scores (as assessed by scenario question types and stand-alone question types) vary according to student gender or ethnicity. First, at the federal level, compliance with the NCLB Act requires an annual report of school progress from each state. The indices used to measure and track school progress are based almost entirely on student achievement scores disaggregated by student ethnicity. Decisions about the distribution of federal aid to school districts are made based on these indices. Thus, it is critical that achievement scores accurately reflect student learning. New information from this study may help accurately assess students' knowledge and skills.

Second, at the state level, students from underrepresented groups fail the Washington Assessment of Student Learning (WASL) at higher rates than other students. Even though Washington State is making progress, the percentage of Native American, African American and Hispanic students in the lowest proficiency level were at least double that of White students on the 2005 8th-grade science WASL. One of the goals of accountability systems in education is to provide students with a fair, equitable means of demonstrating evidence of learning. Thus, it is imperative to examine if scenario question types are more equitable to students of underrepresented groups when compared to traditional stand-alone question types frequently found in large-scale tests.

Third, at the classroom level, teachers strive to determine their students' level of understanding of the science standards known in Washington State as the Essential Academic Learning Requirements (EALRs). To accomplish this, teachers may administer chapter unit tests and tests they have constructed individually or collectively with other teachers. To construct an accurate assessment, it is important for teachers to know the sound principles of assessment including if achievement scores differ by student gender, ethnicity and item format.

Additionally, teachers generally do not resent using valuable classroom time to administer tests, if the test results are helpful to inform classroom instruction (Hattie & Jaeger, 1998). Scenario question types have the potential to be diagnostic because they take a holistic approach with "big picture" science concepts as opposed to a fact-based approach. For example, scenario question types measure science domain knowledge linked to science principles, conditions of application, and reasoning. Often, commercially available tests with stand-alone question types yield little useful information for the classroom (Lohman, 1994). However, WASL scenarios may help bridge the gap between large-scale and formative assessment. Unlike commercially available tests, development of a scenario originates from the essential academic learning requirements (EALRs) and is, therefore, aligned to the state standards. It is not unreasonable to think teachers may be able to pinpoint troublesome EALRs or content areas and change instruction in the classroom if detailed achievement data is accessible.

Finally, assessment developers create carefully designed tasks to evoke evidence of learning. Tasks can be in the form of multiple-choice or short or extended constructed response. Tasks can be set in a context, have one or many steps, be scaffolded, include varying amounts of story text, and can be either abstract in nature or have a practical/real-life application. This research will help inform decisions made by assessment developers by providing a better understanding of how student gender and ethnicity interacts with question type.

CHAPTER 2

LITERATURE REVIEW

The first section of this chapter provides an overview of the study of the measurement of human intelligence in the 20th century. The second part of the chapter examines empirical and theoretical literature on cognition and individual differences. Section three outlines factors that influence student test performance. The final part of this chapter will review advances in cognitive science and educational measurement and their implications to assessment design, concluding with a discussion on the design of the Washington Assessment of Student Learning (WASL) science assessment.

Evolution of the Measurement of Intelligence

The First Half of the Twentieth Century

Assessment and educational measurement is not new in this country. Educational measurement has been discussed in professional communities as far back as the late 1800's. This section of the literature review traces the development of the field of assessment and discusses seminal studies that have shaped popular perceptions of assessment, evaluation, and accountability in education and education policy.

In 1910, John Thorndike published the first scientific tool for measuring educational products (Ayers, 1918). Known as the "Thorndike Scale," this instrument evaluated the merits of handwriting. Thorndike and his students published subsequent standardized tests for reading, language, arithmetic, spelling and drawing (Madaus, Stufflebeam & Scriven, 1983). Thorndike wrote about many concerns that still permeate education today. He acknowledged the challenge of taking complex academic subjects and dividing them into abstract constructs for the purposes of testing. He was concerned these constructs did not provide a complete measurement, but rather “a highly partial and abstract treatment of the product” (Thorndike, 1918, p. 17). Secondly, Thorndike was concerned about the creation and distribution of “mediocre” tests that began to proliferate through the public school system. School administration, teachers, parents, and pupils wanted educational measurement to inform school systems, methods of teaching and promotion of students; in sum, the total educational enterprise. Thorndike cautioned against making inferences from test data that were not confirmed with empirical evidence as well as insisting the accuracy and reliability of test results required multiple measures (i.e., numerous tests over several days with varying test formats and at least two judges to independently rate the test result).

One goal of creating standard, uniform tests was the “removal of gross prejudices on the part of teachers in their own evaluations of certain educational aims and products” (Thorndike, 1918, p.20). Thorndike warned “great care” must be taken to use test results to determine the fate of students, the value of teaching methods, and the achievement of school systems unless the test has been statistically validated and deemed reliable. Fairness of testing was an issue because of “inequities in interest and efforts and inequalities in understanding what the task is” (Thorndike, 1918, p.23).

In 1911, public pressure launched the first public school evaluation (Ayers, 1918). Educational tests evaluated the performance of public schools in New York and firmly established the principle that tests could be used favorably in the measurement of educational processes and products. The Bureau of Research in New York City was established as a separate entity to conduct continuous evaluations to provide information to benefit the school system and by 1913, bureaus had been established in Boston, Detroit, Kansas City, and Oakland. The Binet-Simon tests further engendered public faith and interest in educational measurements (Haertel & Herman, 2005).

In 1917, the United States declared war on Germany and federal attentions turned toward the recruitment of men to support the war effort. Of primacy was the goal of sorting and selecting students for placement in the war. Examinations, called the Army Alpha and Beta, were developed at Stanford University by L. M. Terman, for "mental measurement" of recruits (Terman, 1916, p. xiv.) These tests were a revision of Binet's test and eventually renamed the Stanford-Binet. Once the war was over, Terman recommended using the Stanford-Binet mental test in schools to determine the intelligence quotient (IQ) of the student. He felt intelligence was determined by heredity (Terman, 1916) and buttressed his argument using educational psychology literature published at the time. Terman advocated that children cannot make equal academic progress and recommended placing students in a five-year curriculum track to match the Stanford-Binet IQ. Students with high IQs were given challenging curriculum while students with lower IQs were given less rigorous instruction, a track that would most likely follow students throughout schooling.

Even though the beginning of measurement in education was primarily for the purpose of sorting and selecting students, a new trend emerged in the 1930's led by Dr.

Ralph Tyler from the University of Chicago. Tyler questioned the current practice of determining “educability” of students within the narrowly defined characteristics of the successful student. He criticized general intelligence tests for incorrectly categorizing students and wrote, “it seems clear from such research, that youngsters who do not show up well on intelligence tests, do possess abilities that indicate some skill in solving practical problems” (Tyler, 1948, p. 206). One of the most serious social effects caused by the incorrect categorization of students was that “this practice has tended to deny more adequate educational opportunity to those students who need it most” (Tyler, 1948, p. 211). In sharp contrast to earlier etiological beliefs, Tyler claimed that all persons possess talents, and called for a wider, more robust definition of learning objectives to utilize these different talents. This objectives-based framework lent itself to evaluation of whether or not learning objectives were met and provided feedback for continuous improvement of curriculum. Teachers also participated in the process of curriculum and instructional improvement, refining student measures of learning goals as needed (Haertel & Herman, 2005). In time, this activity was recognized as a “potent method of continued teacher education” (Smith, Tyler, et al., 1942, p. 30). Today’s formative assessment techniques and continuous improvement models have their roots in Tyler’s framework (Haertel & Herman, 2005). To properly measure students’ progress, Tyler called for a variety of test formats and the use of verbal as well as nonverbal materials (Tyler, 1948), recommending that comprehensive records of student performance include descriptions and teachers’ observations, not simply test scores (Haertel & Herman, 2005). These comprehensive records became a means by which to provide school-level accountability, teacher professional development, and clearly articulated learning objectives.

Summary. The first fifty years of public education in the United States can be characterized by behaviorist learning theory, social efficiency, and the hereditarian theory of intelligence. Educational measurement focused on development of tests, such as the Stanford-Binet Intelligence Quotient test, in an attempt to reduce human mental ability to a single, absolute factor that could be reliably measured. Etiological viewpoints were common, allowing test results to influence access to educational opportunities and to predict social roles.

The Second Half of the Twentieth Century

The period from 1950 to 1970 primarily focused on measurement-driven instruction and criterion-referenced testing (Haertel & Herman, 2005). Learning objectives, clustered into cohesive units or "sequences," guided classroom instruction, and end-of-unit tests provided focused, diagnostic feedback on students' progress. In 1956, Benjamin S. Bloom, a student of Dr. Tyler's at the University of Chicago, published *Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain*. Bloom stressed target attainment and in this seminal work, provided a taxonomy for writing test questions by categorizing the level of abstraction of questions frequently found on tests, from lower competency levels (i.e., recall of information) to higher competency levels (i.e., predict and verify evidence). What became important at this time was the comparison of students' progress toward the learning goals, not the comparison of students to each other. This concept became widely known as "criterion-referenced testing" versus "norm-referenced testing" after Glaser's (1963) seminal work:

> Along such a continuum of attainment, a student's score on a criterion-referenced measure provides explicit information as to what the individual can or cannot do. Criterion referenced measures indicate the content of the behavioral repertory, and the correspondence between what an individual does and the underlying continuum of achievement. Measures that assess student achievement in terms of a criterion standard thus provide information as to the degree of competence attained by a particular student, which is independent of reference to the performance of others. (p. 520)

Prior to the Sputnik launch in 1957, funds for educational measurement, evaluation and assessment came primarily from local coffers, foundations, community and professional organizations (Madaus et al., 1983). Federal and state agencies had not yet become involved. However, the launch of the Russian rocket engendered a sense of competitiveness and urgency in the U.S. public. In response, the National Defense Education Act of 1958 was established and provided funds for the development of new mathematics, science and foreign language programs. One of the program outcomes was a new national mathematics and science curriculum from the National Science Foundation (NSF) that would help the U.S. regain competitive advantage. Under the leadership of President Lyndon B. Johnson and President John F. Kennedy, federal monies poured into programs designed to wage "War on Poverty." As funding sources for educational programs shifted from local to federal levels, requirements for the evaluation and assessment of educational programs increased. By 1964, schools were held publicly accountable for federal monies received in such programs as the Elementary and Secondary Education Act (ESEA) Title I. These evaluations made use of standardized test data to evaluate the congruence between stated objectives and outcomes (Madaus et al., 1983).

From 1960 to 1980, cognitive psychologists sought to provide information about learning processes to inform and improve instruction (Lohman, 2000). Researchers focused on the combined study of aptitudes and educational treatments, aptitude treatment intervention (ATI). It was thought ATI would identify the ideal match between student aptitude and education treatment, thereby optimizing individual learning. Researchers used off-the-shelf assessment items that differentiated between students in the hope of isolating measures of different mental aptitudes. After nearly three decades of work, ATI research did not produce the sought-after result. Bond & Glaser (1979) wrote, “no interactions are so well confirmed that they can be used as guides to instruction.” Some researchers have suggested that one reason for this failure was the design of the test items themselves. Off-the-shelf test items are designed to uphold traditional psychometric properties, such as homogenous patterns of response, which are unlikely to provide instructionally relevant information about individual student performance (Lohman, 1994). Further, Resnick (1979) challenged the applicability of performance on test items used in the research studies to the processes of learning. Pelligrino et al. (1999) agreed that the entire approach of ATI seemed removed from classroom learning environments.

The sociopolitical climate of the 1990s changed the focus of assessment from providing information to improve learning and teaching to providing measures of achievement (Pelligrino et al., 1999). One of the criticisms of this shift is the lack of useable results from large-scale summative assessments to inform teaching and learning. As a result, assessment for learning, or formative assessment, is receiving increased attention from education professionals. One way to reconcile differences between summative and formative assessment is to develop state standards that can be disaggregated to a finer level of detail. Ideally, both summative and formative assessments are crafted from the same state standards.

Washington State responded to this challenge in several ways. First, an addendum to the state standards called "Classroom Evidence" provides detailed information about the skill set needed to achieve proficiency. Classroom evidence sections give teachers the opportunity to align lessons to standards, providing continuity between classroom formative assessments and the WASL. Second, OSPI provides student-level WASL achievement data to districts in a timely manner. These achievement data are disaggregated by strand or EALR and provide feedback to a teacher on how well instructional materials and classroom activities are working to achieve the desired learning targets. Although a complete discussion about formative assessment is beyond the scope of this paper, research in formative assessment is providing valuable information to assess student progress, inform classroom teaching and increase student learning and achievement (Rodriguez, 2004; Meisels, Atkins-Burnett, Xue, Bickel, Son, & Nickolson, 2003; Black & Wiliam, 1998).

Perspectives of Learning and Cognition

One approach to organizing viewpoints on learning and cognition is to divide them into four perspectives: trait/differential, behaviorist, cognitive/information-processing, and situative/socio-cultural (Greeno, Pearson, and Schoenfeld, 1996). In the early part of the 20th century, the dominant view of the structure of intelligence could be described as the differential/trait perspective. One assumption of this perspective is that individuals possess stable traits in differing amounts that are measurable. An “effective” assessment task would be able to discriminate between children within and across age ranges (National Research Council, 2001), resulting in different patterns of correlation among test scores. Theories of intelligence based on the differential/trait perspective use these patterns of correlation to infer mental ability. This approach to assessment was developed to evaluate abilities independent from the processes and content of academic learning; however, the design was used to create the “standardized” achievement tests used in the middle part of the 20th century (Marshall, 1995) when behaviorist theories became popular.

The behaviorist perspective of intelligence (Skinner, 1938) asserts that knowledge is organized in a series of stimulus-response associations. These stimulus-response associations serve as the components of skills and can be strengthened by reinforcement or weakened by inattention. Thorndike (1931) defined learning as the process by which one acquires these skills. A few years later, Tyler (1948) organized skills into learning objectives that represented domains of knowledge that could be used to evaluate the extent of students’ learning.

Today, many common achievement tests are influenced by the behaviorist perspective of cognition and assume that mastery of a domain of knowledge can be assessed by analyzing component information, skills, and procedures. Developers of assessments commonly create test items matched to learning objectives and assume that achievement level on such items indicates the level of knowledge in that domain.

In contrast to the behaviorist perspective, the cognitive/information-processing perspective is learner-centered and includes the student as an active processor of information rather than a passive recipient. Mayer (1998) explained his cognitive model of knowledge construction as three separate processes: selecting relevant information, organizing/combining information, and integrating information into existing knowledge structures. When learners actively construct their own meaning, the role of the teacher becomes one of facilitator responsible for activating prior knowledge structures (Driver, 1990), schemas (Marshall, 1995), mental models (White & Frederiksen, 1998), or scripts (Minsky, 1985) and assisting learners to build connections. Mastery of a domain of knowledge, then, is measured not only by how well a student can recall factual information and routine procedures, but by how well a student can integrate and apply the knowledge in different situations. Assessments shaped by the cognitive perspective measure the development of students’ integration of knowledge, procedural skills, and task representation in a domain as students progress toward proficiency (Glaser, Lesgold, and Lajoie, 1987).

The sociocultural/situative perspective arose out of the limitations surrounding a seemingly exclusive focus on individual thinking and learning (NRC, 2001). The sociocultural/situative perspective delineates the role of both the individual and community in developing a learner’s knowledge to participate in the practices of that community. This perspective maintains that knowledge is constructed through social interaction, sharing of culture, and participating in authentic activity. Some researchers have called this situated learning or cognitive apprenticeship (Lave & Wenger, 1991). Based on what is known in cognitive theory, when school becomes the integration and extension of teaching strategies found in the home, students can find meaningful learning. Families and social groups ranging from clubs to classrooms can influence and motivate learners to participate in the practices of the community. To some extent, every assessment includes a measurement of the degree a learner can participate in the practice of test taking (NRC, 2001).

The Study of Cognition and Individual Differences

Empirical and theoretical studies in cognitive psychology support individual differences in performance and learning. The study of individual differences is divided into two approaches: 1) cognitive correlates and 2) cognitive components (Glaser & Pelligrino, 1979). The cognitive components approach attempts to link differences in student test performance to cognitive components such as memory span, spatial visualization, and inductive reasoning. Individual differences have been linked to cognitive components, such as mental models (Glaser & Baxter, 1999; Vosniadou, 1994), schemas (D'Andrade, 1992) and situational differences (Brown, Collins, & Duguid, 1989). The cognitive correlates approach seeks to explain differences in test performance through the mechanics of information processing such as speed, executive routines, and efficiency. Individual differences have been linked to cognitive correlates such as working memory processing speed (Woltz, 2003) and modes of processing (de Ribaupierre & Rieben, 1995; Reuchlin, 1978).

Cognitive Correlates Approach

Ribaupierre and Rieben (1995) postulated that intellectual performance must be understood using a multidimensional approach consisting of modes of processing and situational differences. The model uses processes to distinguish between general and specific variability in an individual's approach to problem solving. The propositional mode of processing is described as analytical, separable by logical rules (Kemler Nelson & Smith, 1989). The propositional mode is well adapted to solving problems requiring formal logic and lends itself to sequential and logico-mathematical (LM) tasks (Ribaupierre & Rieben, 1995). Compared to analogical processes, propositional processes are not as dependent on task context and situational cues. On the other hand, analogical processing is described as global or holistic. This mode is well adapted to solving problems with spatio-temporal relations (i.e., spatial concepts), lending itself to infralogical (IL) tasks. An example of an analogical task is the water-level task, developed by Piaget and Inhelder (1956) to study children's' infralogical cognitive structures. An upright container, presented in an illustration, is approximately one-half filled with liquid. Children are then asked to predict the position of the liquid when the glass is tipped. Findings from many studies suggest that analogical processes are strongly dependent on task context including semantics and illustrations (Bideaud, 1988).

Using these two formats, the authors developed tasks to study the different processes (LM or IL) students evoked when solving problems. They hypothesized that situational differences, as captured by varying representation format, could be one determinant of mode of processing used by the student to solve problems. To test their hypothesis, an assessment was created that included eight Piagetian tasks: 2 LM and 6 IL tasks consisting of 38 items in total. The items spanned different subject domains such as quantification of probabilities, conservation of numbers, and volume. Study participants ranged between 6 to12 years of age. In the first phase, 154 children (22 in each age group) were given the test. In the second phase, three years later, all but the 12-year-old group were given the test a second time to determine if their preferred mode of processing had changed. The authors' analysis confirmed that the LM-IL distinction is significant in understanding individual and situational variability in task performance. Though both processing modes are viewed as developmentally equivalent, results of this study showed certain individuals preferred one type of processing mode over the other.

Given that both processing modes are present in all individuals (Reuchlin, 1978) but vary according to accessibility and evokability, it is possible this could contribute to differences in scholastic achievement (Ribaupierre and Rieben, 1995). On assessments, item formats requiring propositional processing are found more frequently than item formats requiring analogical processing. This may disadvantage students who prefer analogical processing (Globerson, 1989). Further, analogical processing is more prevalent in younger children while propositional processing develops as children grow older (Bideaud, 1988).

Cognitive Components Approach

Knowledge is arranged in conceptual structures referred to as schemas. Schemas are used to interpret and make sense of the world (D'Andrade, 1992). Rumelhart (1980) described schemas using four analogies. First, schemas are like plays and consist of characters, settings, and actions. Schemas are also like theories, filling in missing information and helping to interpret events. Third, schemas are like computer programs, activating particular scripts as information is entered. Finally, schemas are like parsers, decomposing and rearranging information into cognitive structures.

Evidence suggests that schema-based processing is evoked when students solve problems (Marshall, 1995; Cheng & Holyoak, 1985). In other words, the way students approach and solve math and science assessment items depends on the structure of their mental schemas. Further, differential problem-solving schemas are found between different cultures (Vosniadou, 1994; Lopez 1997).

Schemas are organized into mental models that guide and manage the process of problem solving. Mental models that guide reasoning differ between experts and novices (Margolis, 1990). For example, experts tend to use domain-based strategies (Marshall, 1995) to solve problems by recognizing meaningful patterns and matching them to problems encountered before (Margolis, 1990). Extensive stores of knowledge are organized into retrievable well-connected schemas in a way that is closely linked to the problem context (Chi, Feltovich & Glaser, 1981; Larkin, McDermott, Simon & Simon, 1980). On the other hand, novices have less-developed knowledge structures and less-efficient retrieval systems. They are more likely to apply general problem solving strategies by breaking problems into smaller, discrete pieces thereby not seeing the "big picture" (Larkin et al., 1980).

Lopez et al. (1997) sought to identify some universal (cross-cultural) and culture-specific taxonomy schemas and mental models. They compared folk biological taxonomies and inductive reasoning used by industrialized Americans and traditional Itzaj-Mayans. In each culture, the pattern of correlations among participants was strong, indicating a common cultural belief or consensus. Results of the study indicated that Americans made diversity-based inductions whereas Itzaj-Mayans made ecologically based inductions. For example, Americans sorted animals by size while Itzaj-Mayans sorted animals by habitat. Each culture used different patterns of reasoning influenced by different internal categorization of animals, supporting the general hypothesis that categories are cultural projections of the mind (Murphy & Medin, 1985). Thus, the researchers concluded that cultural beliefs influence how science problems are approached and solved.

Cultural Implications for Assessment

Within Western cultural schemas and mental models, then, a child would learn the communication and language idioms that would prepare him or her for the traditional American school experience. Western cultures typically use explicit, direct communication with an emphasis on *content* meaning found in words, while other cultures, such as Native American, emphasize implicit, indirect communication with an emphasis on *context* meaning found around words (Swisher & Deyhle, 1987). Western cultures typically emphasize competition, problem solving and promptness while Native

American and Hispanic cultures emphasize community, acceptance of life's difficulties, and relaxed social experiences (Dejong, n.d.).

The school environment is of particular concern when a students' natal culture is not the dominant European culture of the United States. Curriculum, teaching strategies, and assessments in the United States have been developed primarily with the mainstream European culture in mind and students from other cultures may be disadvantaged and inadvertently penalized on assessments (Solano-Flores, 2002; Delpit, 1988; Hidalgo, Bright, Siu, Swap, & Epstein, 1995).

Current misperceptions in the educational community suggest that children of underrepresented cultural groups operate under a deficit model perpetrated by their cultural upbringing. Henderson and Berla (1994) state "the problem lies not in a lack of preparation for learning but instead in the mismatch between the preparation provided by the home and that which is expected in the school" (p. 149). This mismatch may be perceived as a deficit of the culture; however, Solano-Flores (2002) suggested that the focus should be placed on the deficit of the tests themselves.

Because test development is influenced by schemas, models, and meanings of Western culture, some researchers question the validity of standardized tests for underrepresented cultural groups (Solano-Flores, 2002; Greenfield, 1998). In one such study, Solano-Flores (2002) presented students with two test items from the National Assessment of Educational Progress (NAEP) tests. The sample population included 101 Micronesian students from the Commonwealth of the Mariana Islands (CNMI), 99 immigrant Latino students from rural Washington, and 100 rural Alaskan Yup'ik students. After reading the item, the students were interviewed to determine their understanding of the question. Their responses were categorized into either text/basic skills, context/personal experience, or content/concepts. This organizational system was created to help understand the differences in how students think about and interpret a test question. Most (67%) of the Micronesian students interpreted the question using textual information provided by the item whereas most (50%) Yup'ik students interpreted the question using perceived concepts addressed by the item. Latino students were equally divided between textual, personal experience, and content approaches. These differences in interpretation between ethnic groups may evoke varying modes of processing used for reasoning which affect achievement scores. Further, scenario question formats may favor groups that rely on personal experience to interpret questions while groups that rely on textual and content information may be more sensitive to semantics or embedded illustrations.

Greenfield (1998) worked with unschooled Wolof children in Senegal to test their conservation of numbers. She used the Cambridge method to question the children and to have them justify their judgment answers. The question began with, "Why do you think…", which solicited silence from the children. Only when the question was rephrased as, "Why is….", did the children respond. It is noted that their justifications were as articulate as those solicited by Piaget in the Geneva children. Had this been a high-stakes test, these children would have been penalized for their different epistemology and viewed as unable to reason when the real issue was one of semantics.

Hymes (as cited in Koelsch & Trumbell, 1996) observed that community norms define ways of interpretation:

> If we are to understand what children from a community are saying, and how they hear what we say to them, we must come to be able to recognize more than the language of what is being said. We must recognize how the community norms of interpretation are embodied in speech. (p. 279)

Wittgenstein (1963) suggested that native linguistic patterns resulting in differing organizations of thought might be partially responsible for interpretation differences.

This sampling of research suggests that culturally diverse groups interpret test questions differently. Item contextual factors play a critical role that elicit different responses within ethnic and gender groups. Further, differing cultural schemas (D'Andrade, 1992; Lopez et al., 1997) and mental models (Vosniadou, 1994) provide differential guides of interpretation and inductive reasoning. Students' varying cultural and linguistic backgrounds influence how knowledge is structured and accessed in long-term memory.

Factors that Influence Test Performance

Hattie, Jaeger, and Bond (1999) provide a useful framework to examine key factors within educational testing that can influence test performance. In this model, five over-arching areas affect student test performance: conceptual models of measurement, test and item development, test administration, test use, and test evaluation. In turn, key factors within these areas are discussed. While some assessment literature reports student achievement disaggregated by socioeconomic status (SES), other studies acknowledge the multidimensional nature of student achievement and consider economic, cultural and linguistic diversity together rather than separately (Solano-Flores & Trumbull, 2003; Gay, 2000). This paper will not separately address the effects of economic factors on assessment but rather consider them together with ethnic, cultural and linguistic factors.

Conceptual Model of Measurement

The principled approach to the development of psychometric tests begins with a conceptual model of measurement. Eighty years ago, the conceptual model of measurement was classical test theory and the resulting test scores were both sample- and item-dependent (Loevinger, 1947). In other words, test scales were created from pilot data compiled from a test administered to a sample of students. Item difficulty values were assigned using the proportion of correct responses to the number of students in the sample. This practice produced calibrated test scales that varied depending on the ability of the student sample and of the test items selected (Wright & Stone, 1979). Wright & Stone (1979) illustrated how commonly used percentile ranks and standard scores were dependent on the original calibration sample and items:

> “If my performance put me at the eightieth percentile among college men, I would know where I stood. Or would I? The same score would also put me at the eighty-fifth percentile among college women, at the ninetieth percentile among high school seniors, and above the ninety-ninth percentile among high school juniors. My ability depended not only on which items I took but on who I was and the company I kept!” p. xi

Thorndike (1926), Thurstone (1927) and other researchers acknowledged the need for an absolute scale “on which zero will represent (none) of the ability in question, and 1, 2, 3, 4, and so on will represent amounts by a constant difference” (Thorndike, 1926, p. 4). Indeed, this absolute scale was needed but such a scale would not be devised until a

Danish mathematician, Georg Rasch, introduced his model in a paper at the Berkeley Symposium on Mathematical Statistics in the spring of 1961 (Loevinger, 1965). Rausch (1961) laid the foundation for a new psychometric by using algebra to derive a probabilistic model independent of the sample of test takers and of the items. The probability that a person will obtain a correct response on an item is a product solely of the person's ability and the item difficulty, not the ability distribution of the student sample or specific set of items. Rausch's model is superior to classical test theory for another very important reason. Because an examinee's ability is not defined in terms of a particular test, Rausch's model made possible longitudinal analysis of students' progress.

Most large-scale testing programs now use a measurement model called Item Response Theory (IRT). There are three IRT models, the first of which is Rausch's one-parameter (item difficulty) logistic model. Other logistic models now include additional parameters used to quantify item discrimination (positive and negative) and guessing, commonly used in SAT and ACT assessments. IRT, however, does not isolate the source of difficulty, which can be different for each student that incorrectly answered the item. Item difficulty can vary by gender (White & Fredericksen, 1998) as well as ethnicity (Kupermintz, Le & Snow, 1999) but IRT does not differentiate between different aspects of item difficulty. For example, one student may be having difficulty with reading comprehension while another has difficulty with the particular science principle being tested. Some researchers are experimenting with the addition of parameters that can be used for detecting whether the source of difficulty is due to cognitive processing, language differences (Erikan, 1998) or solution strategies (Lane, Wang, and Magone, 1996).

Ideally, educational measurement should be integrated with principles of cognitive psychology. Biggs and Collins (1982, 1989) developed the Structure of the Observed Learning Outcome (SOLO) model that allows for personal variation in experience by acknowledging that different knowledge structures can lead to the same observed (test) behavior. The model makes both content knowledge and cognitive processing structures transparent, allowing the model taxonomy to identify the current stage at which the student is operating.

<u>Summary.</u> Most large-scale testing programs now use a conceptual model called item response theory (IRT) because IRT eliminated most of the shortcomings of classical test theory. However, IRT ignores many advances in the area of cognitive processing (Hattie et al., 1999). For example, cognitive studies have shown that complex processes underlie school achievement influenced by the organization of a students' knowledge, such as schemas (D'Andrade, 1992; Lopez et al., 1997) and mental models (Vosniadou, 1994; Solano-Flores, 2002). Several researchers have created an integrated measurement model (Mislevy 1996; Fischer 1997) trying to reduce cultural differentials, but these cultural differences continue to be reflected in test performance (Cole 1996) and can become cultural handicaps (Anastasi & Urbina, 1997). Hattie et al. (1999) urged assessment scholars to apply more cognitive processing measurement models rather than static measurement models (e.g. IRT) to devise test items.

Test and Item Development

Once the conceptual model of measurement is chosen, test and item development begins with the goal of identifying constructs and choosing item response formats, scoring rubrics, and test assembly models. Task construction should be tightly linked to the chosen model of cognition and learning (NRC, 2001). Researchers report that individual differences, such as gender and ethnicity, interact with item response format on standardized tests resulting in achievement differences (Lawrenz, Huffman, & Welch, 2000). One specific concern is that females and non-whites tend to score much lower than white males on science and mathematics multiple-choice tests (Jones, Mullis, Raizen, Weiss & Weston, 1992). One possible explanation for this gender difference is that multiple-choice item response formats favor students who guess (Rowley, 1974) and male students tend to guess more than female students (Hanna, 1986). However, this explanation has been called into question because the research indicates that females tend to receive higher scores than males on multiple-choice items in the domain of reading and language (Langer, Applebee, Mullis & Foertsch, 1990).

Klein et al. (1997) further investigated gender differences by examining characteristics of the question. For example, girls tended to do better than boys on questions requiring interpretation from observations whereas boys tended to do better than girls on questions requiring predictions. Lawrenz et al. (2000) used proficiency level to stratify girls' and boys' performance on multiple-choice, open-ended, and hands-on investigations. He found no significant difference between genders, but found proficiency level to be significant in all item response formats.

It has been speculated that students who score poorly on multiple-choice standardized tests may show improved achievement in performance assessments. Test items on performance assessments do not simply involve a recall of facts (Solano-Flores & Shavelson, 1997) but prompt test takers to write, draw, and display data thereby emphasizing the process by which students generate solutions. It should be noted that the term “performance assessment” initially referred to a task with a “hands-on” component; however, today the term is broadly used to refer to open-ended item response formats other than multiple-choice (Sugrue, 1997).

The Chinle Portfolio Project on the Navajo Nation is one example of a performance-like assessment that successfully used qualitative information in portfolios to satisfy state requirements while honoring cultural ways of knowing is (Koelsch, 1994). Cross-cultural portfolios use “task shells” that specify domains of performance for portfolio tasks but leave the content flexible. When assessment tasks are structured in this manner, students can construct their own meaning by activating prior knowledge structures and schemas while assimilating new information (Driver, 1990; Bransford & Stein, 1984). Indeed, students of underrepresented groups tend to achieve higher scores on alternative assessment formats such as performance-based assessments when compared to traditional multiple-choice and short answer formats (Stecher, Klein, Solano-Flores, McCaffrey, Robyn, Shavelson, & Haertel, 2000).

Research suggests that item context has an effect on comprehension. Pictures, topics, titles (Schallert, 1976) and story passages (Bilsky, Blachman, Chi, Mui, & Winter, 1986) can facilitate comprehension strategies. For example, Bilsky et al. (1986) found that the accuracy of inferences drawn by students improved when targeted problem sets were embedded in a series of story passages. Solano-Flores (2002) notes that minority cultures emphasize contextualized knowledge rather than item independence and subject fragmentation, as is common in multiple-choice tests. The embedded context format may be more accessible to underrepresented groups, allowing easier interpretability and cueing initial thoughts on how to solve the problem (Zumbach & Reimann, 2002).

Summary. When looking at a synthesis of the literature, it is difficult to say with certainty how question format influences student performance. Variation between ethnic and gender groups is even less well understood. Some research suggests that certain formats are more conducive to higher academic achievement in underrepresented populations. However, the reasons for increased performance within underrepresented groups are not well documented in the literature at this time.

Test Evaluation

Many topics fall under the category of test evaluation. The topics of validity and bias are particularly salient to include in a discussion on high-stakes achievement tests and student subgroups. Gliner & Morgan (2000) defined validity as "an evaluation of scores on a particular test and how these scores will be interpreted" (p. 319). Traditional measurements of validity consist of content, criterion, and construct validity. Bias is also considered a type of differential validity (Cole & Moss, 1989).

Researchers have further examined the meaning of validity from a socio-cultural perspective and have challenged the accuracy of inferences made about test score meaning. Solano-Flores and Nelson-Barber (2001) propose that cultural validity, as a form of test validity, should be incorporated into assessment practices. Solano-Flores (2002) defined cultural validity as:

> the effectiveness with which an assessment addresses the socio-cultural influences that shape student thinking and the ways in which students make sense of items and respond to them. These socio-cultural influences include the sets of values, beliefs, experiences, communication patterns, teaching and learning styles, and epistemologies inherent to students' cultural backgrounds, and the socioeconomic conditions prevailing in their cultural groups. (p.1)

Unlike other approaches intended to ensure test equity and fairness, an approach based on cultural validity proactively incorporates cultural diversity throughout the process of assessment development rather than treating it as a source of score variance between cultural groups at the end of the process. Harris, Schoen, and Lee (1986) proposed that the social and cultural schemata students bring to the classroom play an important role in the learner's ability to understand and remember new information. Anchoring knowledge and assessment to culture will not only increase the likelihood of the new material being recalled (Lockhart & Craik, 1990) but will also validate the authenticity of the activity as socially and culturally relevant (Harris, Schoen, and Lee, 1986).

Cole and Moss (1989) defined bias as "differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers" (p. 205). When language affects understanding, then various types of bias can be introduced into assessments (Van de Vivjer & Poortinga, 1997). Occasionally, assessments are translated from English to the test taker's native language; however, translation can become troublesome because, at times, words that have equivalent meanings cannot be found between languages. Test translations are often done in short time frames and are poorly done (Van de Vivjer & Poortinga, 1997), resulting in questionable equivalence to the original test (Solano-Flores, 2002). Poor item translation, complexity of item wording and unfamiliarity of an item can introduce item bias into assessments (Van de Vivjer & Poortinga, 1997) and lead to test inaccuracies (Solano-Flores & Nelson-Barber, 2001).

In addition, students may suffer from unclear instruction from the proctor regarding test directions and how they will be assessed. For example, some cultural groups value brevity in a response (Solano-Flores & Nelson-Barber, 2001). Cross-cultural differences in response styles such as this can be responsible for method bias in the testing instrument (Van de Vivjer & Poortinga, 1997). Test takers ought to understand that their answers may need to be lengthened to avoid an inaccurate evaluation of the amount of knowledge they know. Furthermore, too few questions in the test covering broad constructs can lead to poor test performance and provides an inaccurate representation of students' actual knowledge and comprehension.

Culturally competent test evaluation is complex. Even in ideal performance-based, instructionally embedded assessments, many researchers question how test evaluation can be free of bias when students and teachers do not share the same background or understand students' home culture (Espinosa, 2005; Sanchez & Brisk, 2004). The Federal Public Law 102-119 states that evaluations of students must be free of all racial and cultural discrimination and be conducted in the students' native language when possible (Lynch & Hanson, 2004) but implementation of this public law is problematic (Espinosa, 2005). Sanchez & Brisk (as cited in Espinosa, 2005) reported:

> Most teachers have not been trained to conduct nondiscriminatory assessments with children from culturally and linguistically diverse backgrounds; many do not speak the child's native language and are not familiar with the home culture and; most teachers lack knowledge of the psychometric characteristics of tests and therefore cannot make informed judgments about the appropriateness of specific tests when their students are from linguistically diverse backgrounds. (p. 848)

Test Administration

A key factor that can influence student test performance but can be somewhat mitigated during test administration is limited English proficiency (LEP). During administration of state standardized tests, every possible attempt is made to control extraneous experiences and environmental variables for test-takers. District assessment coordinators and test proctors strive for uniformity of test settings, procedures, and times; however, one exception is test administration with accommodations. Accommodations are modifications in normal test administration procedures and are specifically designed for students with testing handicaps such as Limited English Proficiency (LEP). Accommodations for LEP students frequently include extra assessment time, oral directions in native language, visuals, glossaries in both native and English language, and reading aloud of questions with explanation. Accommodations are important for a more accurate picture of subject-matter knowledge and cognitive abilities and, in theory, can offset distortions in scores caused by the testing disability (McDonnell, McLaughlin, & Morison, 1997); however, matching the appropriate level of accommodation to the student is problematic (Hattie et al., 1999). These accommodations are not without criticism. Solano-Flores and Nelson-Barber (2000) state these approaches are not enough to ensure equity and fairness in testing.

Under federal pressures of accountability, school districts have been tempted to exclude students with Limited English Proficiency (LEP) from standardized testing requirements (McGill-Franzen & Allington, 1993). Shepard, Taylor, and Betebenner (1998) stated, “Exclusion [of English-language learners] removes these children from the accountability system and denies their rights to be full beneficiaries of educational reform efforts” (p. 1). Some state and local school districts have also tended to concentrate efforts on raising achievement scores of “bubble” students just below the cut score of the proficient level. For these reasons, the effects of accommodation in test administration is receiving renewed attention especially regarding the testing medium, time limits and test content (Hattie et al., 1999).

Test Use

Messick (1989) is often cited as the first researcher to expand the notion of test validity to include test use. In his seminal work, he stated, “…what is to be validated is not the test or observation device, but the inferences derived from test scores or other indicators; inferences about score meaning or interpretation and about the implications for action that the interpretation entails” (p. 13). Messick (1994) urged assessment developers to clearly define constructs and detailed scoring rubrics prior to writing assessment items so that correct inferences about students’ knowledge can be made.

Test scores are administratively used for public reports of student achievement, program evaluations and federal accountability requirements (NCLB, 2002). In the first case, seemingly simple matrices (or “league tables”) are used to report student achievement to the public but are criticized as insufficient reporting instruments (Hattie et al., 1999). There is a tendency for the public to use these data to make longitudinal and subgroup comparisons. For example, researchers attempted to link student state achievement scores to the National Assessment of Educational Progress (NAEP) without success due to different units of measure and different test instruments across subgroups (Feuer, Holland, Bertenthal, Hemphill, & Green, 1998). Although the public may view such a comparison as having face validity, the comparisons are highly questionable (Kane & Staiger, 2001; Sicoly, 2002). School districts often report longitudinal data by grade level as evidence of improvement, but cohorts and norm groups may have differing abilities each year, resulting in an inaccurate picture of improvement.

Second, tests are used to evaluate program and curricular effectiveness and to evaluate the adequacy with which essential academic learning requirements are taught. Individual schools primarily have the responsibility of choosing a curriculum that sufficiently covers academic learning requirements. For the diverse classroom, curriculum that is responsive to both constructivist and sociocultural principles of learning has been demonstrated to increase student knowledge among Native American students (Tharp, 1982); however, standardized tests may not reveal an accurate picture of student knowledge. The degree of overlap between a test and curricular content is a strong predictor of student achievement (Cooley and Leinhard, 1980; Ruiz-Primo, Shavelson, Hamilton & Klein, 2002). This implies that students instructed with culturally relevant classroom materials may be at a disadvantage and perform poorly if the testing instrument does not overlap with the classroom materials (Leinhart & Seewald, 1981).

Lastly, standardized tests are used to meet federal reporting requirements. Unfortunately, the lack of consistent reporting methods causes errors (Hattie et al., 1999) and abundant measurement misunderstandings (Frisbie, 2005), which can aggravate adverse impacts to diverse student groups. Data complexities and inconsistencies must be acknowledged and explored before data are used to inform education policy and practice.

Advances in Cognitive Science and Educational Measurement

Matlin (2005) defines “cognition” as “mental activity [that] describes the acquisition, storage, transformation, and use of knowledge" (p.2). Given that cognition cannot be measured directly, one must develop methodologies that can indirectly measure the process of acquiring knowledge. The field of cognitive psychology is rich with methodologies that can indirectly approximate cognitive constructs. Methods such as structured interviews, response patterns across a set of items, response latencies, patterns of errors, written or oral explanations, and task performance are regularly used to investigate theories of learning and memory (Sugrue, 1997). Thirty years ago, Glaser et al. (1987) challenged education professionals to take and adapt these methods for use in the field of educational assessment. Adaptation, however, was difficult because these methods were not developed with traditional psychometric properties in mind, such as item calibration and item fit, which are methods for evaluating plausibility of student responses (Wright & Stone, 1979). At that time, the field of educational measurement did not offer a valid, reliable technique to assess students' competencies using these methodologies (Broadfoot & Black, 2004).

One method appeared particularly promising for science educators to measure science domain knowledge. Performance assessments had the potential to not only assess students' cognitive problem-solving abilities in the science domain, but to also provide a desirable authentic "hands-on" context. In the 1980s, assessment designs began to include performance assessments although their construct validity was questioned (Royer et al., 1993) because of low correlations between the performance-based tasks and multiple-choice tasks, suggesting different aspects of science achievement had been measured (Baxter, Shavelson, Goldman, & Pine, 1992). Noted researchers in the field of educational measurement focused on ways to devise a valid, reliable method of scoring performance assessments. In one such study, the authors evaluated a procedure-based scoring system that not only included a "hands-on" task but also a surrogate notebook completed by the test taker (Baxter et al., 1992). Correlation coefficients between the test taker's notebook and the multiple-choice items increased and students' observed performance scores approximated multiple-choice scores.

Over time, the term "performance assessment" evolved to mean a series of constructed-response items informed by general principles of performance task design (Goldberg & Roswell, 2001). These items typically prompt students to write, draw, and display data in their response. For example, the NAEP science test includes a performance assessment task and surrogate notebook for students' entries. A student's competency is not evaluated from the performance task but rather from entries made in the notebook (NRC, 2001). This departure from traditional multiple-choice items provides a foundation for assessment that may engender improved authenticity and cultural validity in large-scale science assessments.

The Birth of the National Research Council's Model of Assessment

In 2001, the National Research Council convened The Committee on the Foundations of Assessment. Their goal was to synthesize current research on cognition and educational measurement and suggest promising assessment strategies, particularly in the domains of mathematics and science. Multiple committees worked together for three years and produced a seminal book entitled *Knowing What Students Know: The Science and Design of Educational Assessment*. In the book, the authors stressed that high quality assessment design that elicits evidence of learning on intended constructs is not an art, but a principled approach to valid inferences of students' achievement. To guide this principled approach, they introduced a model of assessment called "The Assessment Triangle." Each vertex of the assessment triangle represents one of the three key elements found in every assessment: 1) the model of cognition, 2) observation and 3) interpretation.

The Assessment Triangle – Model of Cognition

The first vertex, model of cognition, is a theory about the nature of learning and the development of competence in a subject domain. Each model of cognition has three key features. First, it is based on empirical studies of learners in the domain. The mathematics and science domain are rich with research on how most students progress in their understanding of particular concepts. In the science domain, detailed models of cognition describe how students would typically represent knowledge of science principles and concepts. For example, Siegler (1976) studied students' ideas about torque in balance-scale problems. Common patterns of ideas emerged and were used to develop a model of cognition outlining the different rules students use to solve balance-scale problems. In another example, an experienced physics teacher noticed how learners attempt to make sense of gravity and its effects on objects in a fluid medium (Minstrell, 2000). He grouped these ideas into clusters and arranged the clusters on a continuum from correct to problematic understandings, illustrated in Appendix A. This model of cognition is diagnostic and is useful to guide classroom instruction and provide feedback to students as well as to design assessments.

The second key feature of the model of cognition is that it differentiates between novice and expert performers in the domain. The novice-expert paradigm is based on the work of Chi, Glaser and colleagues (Chi, Feltovich, & Glaser, 1981; Chi, Glaser, & Farr, 1988; Glaser, 1992; Glaser, Raghavan, & Baxter, 1992) who studied differences between the way experts and novices organize, retrieve, and apply their knowledge. They found that experts encoded knowledge in large sections tightly integrated with principles and conditions of use. Experts' recognition of a particular principle in a task item signaled application, and large subsets of knowledge were easily retrieved. Novices, on the other hand, encoded knowledge in a more fractured structure and did not readily recognize principles or conditions of use. These distinctions help categorize the differences in cognitive components used by experts and novices in problem solving.

Additional research on expert/novices theories is particularly noteworthy. Smith and Good (1984) suggested that novices are often good problem solvers but use different methods than experts. Smith (1991) refined the model based on cognitive components exhibited by good problem solvers who are not yet performing at an expert level. He stated that the expert-novice paradigm is based too heavily on a model of expert problem solving and that the goal of education is not to produce experts but to produce good problem solvers. Smith's model is similar to the model of cognition used to develop the science WASL.

Over the last two decades, expert/novice theories and research have informed models of cognition but were largely derived from studies of experts from areas outside of the academic domain such as chess (e.g. Chase & Simon, 1973), sports and dance (e.g. Allard & Starkes, 1991). One criticism is the difficulty of translating this body of research into useable educational practice (Hatano & Oura, 2003). Like Smith (1991) Alexander (2003, 1997) sought to translate expert/novice research into helpful information within academic settings. Alexander (2003) updated the expert/novice paradigm espoused by the NRC (2001) with a model created exclusively for use within educational practice. Her Model of Domain Learning (MDL) has several fundamental differences compared to traditional expert/novice models.

MDL focuses on learning in academic settings such as astrophysics, human biology and educational psychology rather than non-academic settings. The author suggested that an academic focus is critical for transferability of research to the classroom because of unique characteristics of academic settings (Shulman & Quinlan,

1996) such as socio-cultural and motivational influences (Pintrich, Marx & Boyle, 1993). MDL does not draw as sharp a contrast between an expert and novice as do traditional models. The distinction between a student who is an expert and a student who is a novice is not as central to the Model of Domain Learning (MDL) as the transformations that occur between these two points (Alexander, 2003).

The third feature of the model of cognition is that it incorporates a sociocultural perspective by acknowledging the diversity of learners. This perspective is responsive to differences in age, culture, gender, and language and describes multiple paths to proficiency. One way to develop this perspective is to study educational research on groups of students that reflect the diversity of the student population. Another way may be adopt and align multiple curricula to state standards to serve the wide variety of needs of ethnic and linguistic groups.

An ideal model of cognition is flexible enough to meet the needs of both classroom and large-scale assessments. For example, the model could be used to design a large-scale assessment by including attributes central to the understanding of the science domain. On the other hand, classroom assessments could also be designed by including finer-grained attributes that provide an insight into misconceptions and incomplete knowledge and suggest the next steps in instruction. Each assessment could represent a subset of the model and resulting performance data could be aggregated or disaggregated, depending on the purpose of the assessment. For example, student profiles of performance could be aggregated to report summary information to administrators or disaggregated to report detailed information to teachers.

The Assessment Triangle - Observation

The second vertex in the assessment model is observation: a set of beliefs about types of evidence from which inferences about learning are drawn. The word "evidence" refers to students' test responses. There are different beliefs or theories about which types of evidence demonstrate competencies. These beliefs drive the kinds of assessment data sought and influence decisions about tasks, item formats, and test assembly. For example, Smith's (1991) model of cognition purported that one trait of a good problem solver is ability to apply science knowledge to solve a problem. A released eighth-grade science WASL scenario entitled "In The Doghouse," designed to evoke evidence for such a trait, is included in Appendix B. Optimally, observation should be linked to the model of cognition in such a way that strategies and approaches students use to solve problems are clear.

The Assessment Triangle - Interpretation

The third vertex, interpretation, is a process of making a determination about students' competencies based on students' test responses (observations). It is a process of reasoning from the evidence generated by the observational situations back to the competencies listed in the model of cognition. The interpretation of student responses is achieved with a measurement model that should be consistent with the underlying model of cognition (NRC, 2001); however, current federal and state school accountability requirements constrain measurement model choices. The measurement models used are those that can draw conclusions about large amounts of data and give precise information on measurement error associated with the conclusions (NRC, 2001), a priority in a high-stakes testing environment. Today, on large-scale tests, item response theory is the measurement model most commonly used to translate student responses into achievement scores.

Summary. The assessment triangle symbolizes the desired link between each of the vertices: model of cognition, observation, and interpretation. The triangle illustrates the connection and interdependence of each element, which drives the design of high quality assessments. The authors suggested that a well-designed effective assessment begins with a clearly stated model of cognition. From explicit competencies stated in the model of cognition, carefully designed tasks or observational situations can evoke evidence of learning for the intended constructs. Interpretation, the process of reasoning from students' test responses, is most commonly achieved by using IRT. In the following section, important features of the ideal assessment model explicated above are compared to features of the 2005 eighth grade science WASL.

The Design of the 2005 Eighth-Grade Science WASL

Washington State science standards are organized into strands. The three strands used to unify and categorize ideas of science are Systems, Inquiry, and Application. Each strand includes both Essential Academic Learning Requirements (EALRs) and Grade Level Expectations (GLEs) in the science domain. The Systems Strand describes specific domain knowledge such as concepts and principles. The Inquiry Strand describes components of problem solving needed to investigate the concepts and processes of nature. Lastly, the Application Strand describes the cognitive functions needed to understand how to apply knowledge.

Science WASL - Model of Cognition

The EALRs and GLEs in each strand reflect the model of cognition from which assessments are designed and created. The learning requirements and expectations in each strand reflect a developmental model of cognition depicting increasingly sophisticated stages of understanding as students move through elementary, middle, and finally high school. Within the GLEs, sections entitled “evidence of learning” describe students’ advancement in domain knowledge and strategic processing as they move from a level of acclimation to one of proficiency. These progressions from acclimation to proficiency most closely resemble the Model of Domain Learning espoused by Alexander (2003) rather than traditional novice/expert models that draw sharp distinctions between the two categories. As domain knowledge matures qualitatively and increases quantitatively, the level of strategic processing moves from surface-level to deep-processing (Alexander, 2003) as students move from Kindergarten to twelfth grade. An example of the progression of understanding within the science Inquiry Essential Academic Learning Requirement (EALR) IN02 2.1.2, Planning and Conducting Safe Investigations, is provided in Appendix C.

Currently, the GLEs are not at a level of detail to diagnose students’ misunderstandings or main obstacles to learning. Indeed, they are not written with intent to be diagnostic (R. Beven, personal communication, December 18, 2006). However, GLEs can be integrated into instruction and classroom assessment, which is one feature of an ideal model of cognition (NRC, 2001). The GLEs can be used to guide instructional and curricular decisions as well as to provide a benchmark for formative and summative assessment activities. For example, teachers can use the GLEs to craft questions to initiate class discussion, develop pre- and post-instructional classroom assessments, and provide rubrics for teacher- and student-led evaluation of student work. Additionally, Powerful Classroom Assessments (PCAs) are scenarios developed by OSPI to support formative assessment. PCAs are developed from released WASL scenarios or from curricular materials. They consist of a scenario, scenario items, and a scoring guide. Each short and extended constructed response item has a set of ten student responses with annotations describing how each response was scored. Teacher and student participation in classroom use of PCAs has been demonstrated to have a positive effect on student learning and achievement (OSPI, 2006).

Science WASL - Observation

The design of observational situations, or test items, begins with the model of cognition. General Learning Expectations (GLEs) and evidence of learning are used to develop test items within a scenario. Science WASL tasks are constructed from these well-developed descriptions of understanding; this practice helps ensure the intended construct is measured and the chance of measuring unintended aspects, such as writing mechanics, language proficiency or sentence construction are minimized (Messick, 1994).

The scenario format offers a way to design items that measure domain knowledge (i.e., Systems Strand – Energy Transfer), procedural knowledge (i.e., Inquiry Strand –

Planning an Investigation), and strategic processing components (i.e., Application Strand – Evaluating Potential Solutions). The well-contextualized items target an understanding of key concepts and related principles linked to conditions of application, a common trait of good problem solvers (Chi, Glaser, & Farr, 1988; Glaser, 1992).

Solving problems implicitly requires multidimensional ability (Embretson, 2004; Alexander, 2003). For example, science problems may require computation, planning strategies, encoding, etc. In keeping with the assessment model, "if performance on measuring tasks is multidimensional, then processes must be designated as construct-relevant or construct-irrelevant. Such designation, in turn, guides item design" (p.59, Embretson, 2004). Tasks within scenarios are multidimensional tasks that provide the unique ability to align content and strategic processing constructs needed to answer questions correctly.

Using several different item formats and several different response types, students are able to use words, pictures, and numbers to convey intellectual abilities and demonstrate evidence of learning. Question stems such as "Write a plan" or "Explain why you choose your approach" are used throughout a scenario. These stems cue test-takers and help make students' thinking clearer compared to simple multiple-choice formats.

Science WASL – Interpretation

Like most large-scale tests, Washington State uses Item Response Theory (IRT) for a measurement model to interpret student performance. IRT takes a probabilistic approach (Wright & Stone, 1979) rather than a cognitive approach to reasoning. A statistical model is developed for each item, characterizing response patterns most likely to occur from students at varying levels of competency.

According to cognitive theory, the ability to solve the science WASL scenario questions involves a complex interaction between cognitive/strategic processing components (Alexander, 2003; Alexander & Judy, 1988), content/domain knowledge (Peverly, 1991), and student affective variables such as interest and beliefs about oneself (Baghetto, 2007). If learning is the process of moving along the continuum of cognitive constructs (e.g. knowledge structure, strategic processing) from novice to expert (Chi et al., 1988), then the method of interpretation would need to be sensitive enough to capture the qualitative and quantitative differences in problem solving between novices and experts. Value points offer a finer granularity for interpreting students' responses compared to score points because each score point is the equivalent of two value points. This finer granularity may provide a way to determine when a students' domain knowledge is sufficient to answer a question correctly but their strategic processing or conditions of applicability are misinformed. Ideally, analyzing why a student answered a question incorrectly rather than simply calculating the percentage of students who answered a question incorrectly would help identify the sources of difficulty students are having.

Summary. The discussions in this chapter explicate some of the critical features of the ideal model of learning, observational situations, and inferences put forth by the National Research Council (2001). These critical features are applied to the 2005 Eighth-Grade Science Washington Assessment of Student Learning. The model of cognition used to develop the science WASL is most representative of the Model of Domain Learning (MDL) espoused by Alexander (2003). The MDL illustrates the transformations that occur as a student progresses from Kindergarten to Grade 12 rather than sharp distinctions between experts and novices found in traditional expert/novice theories. The multidimensional nature of scenarios offers a way to measure students' knowledge of science concepts, principles, application, and strategic processing in an integrated way. Washington State uses the IRT model to interpret test evidence. This model is not ideal because it does not indicate why a student might have given an incorrect response. However, using value points rather than score points provides a finer level of granularity that may be helpful to examine socio-cultural differences that can account for differences in academic performance.

CHAPTER 3

METHODS AND PROCEDURE

This chapter will provide a brief history of Washington State's education reform policy and the unique collaborative process of constructing the science Washington Assessment of Student Learning (WASL). The chapter will also describe the research design, data instrument, and data analysis for this study.

Background of the Study

Historical Overview of WASL Development

In the late 1980s and early 1990s, the Washington Office of the Superintendent of Public Instruction (OSPI) began to receive feedback from the state universities that students were not prepared for college-level classes (R. Houser, personal communication, May 15, 2004). On the average, it took approximately one year of remedial work to bring high school graduates up to the level needed to be successful in college coursework. Local business leaders also sent OSPI the message that graduates were unprepared for adulthood and lacked the job skills they needed. The former State Superintendent responded to these messages by assembling a committee of key policy makers and stakeholders to create a state education reform package. In 1993, the Washington State Legislature approved the reform package and implementation began.

The first step in the reform package called for the creation of a Commission on Student Learning. Membership in this commission was through application, which was reviewed by the governor and a review panel. Next, this commission developed two primary goals: 1) identify subject strands and standards for content areas in which students were deficient and 2) create a criterion-referenced evaluation tool that would measure student learning progress in reading, writing, mathematics and science. Two years elapsed before subject strands, standards, and an evaluation tool were created to the satisfaction of the Commission. The outcomes came to be known as the Essential Academic Learning Requirements (EALRs) and the Washington Assessment of Student Learning (WASL), which together set the foundation for the educational reform program.

The initial goal was to implement the tests in 4th, 7th, and 10th grades. The 4th grade test was administered on a voluntary basis in spring 1997, the 7th grade test was administered in spring 1998 and the 10th grade test was administered in 1999. Following a successful initial implementation of the tests, the Commission on Student Learning was dissolved and the Academic Achievement and Accountability Commission (A+ Commission) was created by the 1999 Legislature to provide continuing oversight of the K-12 educational accountability system. The nine-member commission includes a teacher, a principal, a district superintendent, an attorney, an education researcher, three business individuals, and the State Superintendent. The members were selected by Governor Gary Locke from nominations from the Washington State Senate and House of Representatives and recommendations from statewide groups and the public. The A+ Commission continues to meet with the public and business leaders in public hearings to solicit input on improvements that can be made (R. Houser, personal communication, May 15, 2004).

Washington State was poised to meet the requirements of the federal No Child Left Behind (NCLB) Act, which was launched in 2002. In fact, Washington D.C. used the state of Washington as a model for other states regarding how to build an infrastructure that would effectively support education reform (R. Houser, personal communication, May 15, 2004). This infrastructure includes many stakeholders that are involved in the process of continued improvement of the state education reform policy. Stakeholders' involvement includes the Technical Advisory Committee at state and national level, the State Board of Education, the Academic Achievement and Accountability Commission, and the Office of the Superintendent of Public Instruction. All these groups work as pieces of the puzzle that come together to implement education policy and to develop plans to be taken to the state legislature.

Current Assessment Policy

WASL tests are administered in Grades 2-10 and have varying content focuses in each of the grades (Figure 1). The science WASL is administered in Grades 5, 8, and 10.

GRADE LEVEL									
Content Area	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Reading	X	X	X	X	X	X		X	X
Writing			X	X		X			X
Language Arts					X			X	X
Science				X			X		X
Mathematics					X			X	X

Figure 1. Grade levels and content areas that are the focus of WASL testing (OSPI, 2005).

The Washington State science standards prescribe what all students should know at each of these grade levels. The standards are divided into three sections: Systems, Inquiry, and Application. The "Systems" section consists of content knowledge about properties, structure, and changes in physical, earth, space, and living systems. The "Inquiry" standard describes how to scientifically investigate a system to answer questions. In other words, students are expected to engage in the process of inquiry to discover new knowledge. Once an understanding of the system is achieved, the knowledge can be used to design solutions to human problems dealing with science, technology, and societal issues. This design process is the third science standard called "Application." This holistic approach to science education provides teachers and students with a unified way of thinking (Figure 2).

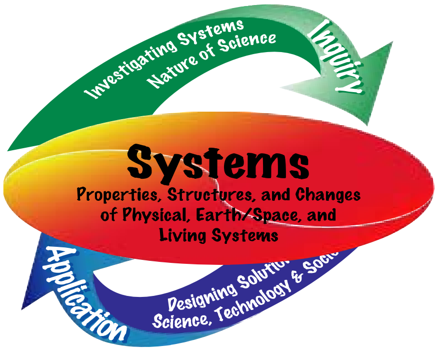

Investigating Systems
Nature of Science
Inquiry
Systems
Properties, Structures, and Changes of Physical, Earth/Space, and Living Systems
Application
Designing Solutio
Science, Technology & Socie

Figure 2. Washington State's science standards are grouped into systems, inquiry, and applications (OSPI, 2004).

Essential Academic Learning Requirements (EALRs) are organized around these three interrelated and interdependent science sections: Systems (What), Inquiry (How), and Application (Why). To facilitate communication of learning expectations to both teachers and students, each EALR is broken down into many Grade Level Expectations

(GLEs). Each GLE includes an “evidence of learning” section with clearly stated expectations of the student within each grade level. This helps inform classroom teachers in designing the next learning opportunity.

Three types of scenarios are developed, modeled after the Systems, Inquiry, and Application sections to measure student understanding of concepts and processes. First, “Systems” scenarios begin with brief descriptions of a system, such as a bean seed growing in a terrarium. Students are asked about the inputs, outputs, and transfers of matter, energy, and/or information in the system described. Often students are then asked to plan a new investigation for a related question in the second type of scenario called “Inquiry.” In their work, students demonstrate their understanding about the attributes of the investigation and the system being investigated. Lastly, “Application” scenarios, recently changed to “Design” scenarios, require students to use the results of the investigation to design a scientific solution to a human problem. Students demonstrate their understanding about their experimental design and how they applied scientific concepts. Multiple types of student evidence (e.g. charts, pictures, text) can be used to demonstrate knowledge. Stand-alone questions are not linked to a scenario and are either multiple-choice or short answer.

Of the 45 items on the 2005 eighth-grade science WASL, five (10%) measured Properties of Systems, six (14%) measured Structure of Systems, seven (16%) measured Changes in Systems, 18 (40%) measured Inquiry in Science, and nine (20%) measured Application of Science.

The Participants

In 2005, 81,690 enrolled eighth grade students completed the eighth grade science WASL. For this analysis, only valid, complete test records will be included. Reasons for invalid test records included cheating, absence, Limited English Proficiency exemption, medical exemption, and not enrolled status. Incomplete records may be missing gender or ethnicity designations. The elimination of incomplete and invalid student records resulted in a data set of 77,692 student records. Females and males were more or less equally represented. The number and percentage of participants by ethnicity are presented in Table 1.

Table 1. Number and Percentage of Participants by Ethnicity.

Ethnicity	*Number of Participants*	*Percentage*
American Indian / Alaska Native	2189	2.8%
Asian	6043	7.7%
Black African American	4324	5.6%
Hispanic	8709	11.2%
White	56,094	72.2%
Hawaiian Pacific Islander	69	< .001%
Multiracial	264	< .005%

Research Design

This type of quantitative study is ex post facto/causal comparative, although there is some disagreement among authors on whether to include the word "causal" (Gliner & Morgan, 2000). Gliner & Morgan (2000) label the approach individual differences or nonexperimental because there are no active independent variables. There are two normally distributed dependent variables: 1) total score on scenario question types and 2) total score on stand-alone question types. There are two independent variables, ethnicity with seven levels (American Indian/Alaska Native, Asian, Black African American, Hispanic, White, Hawaiian Pacific Islander, and Multiracial) and gender with two levels (Female, Male).

Data Instrument

Requirements of the science WASL are as follows: 1) the items must align to a valid construct to be measured (as defined in the EALRs), 2) the items discriminate correctly based on test takers' ability (measured by point biserial values), 3) the measures are not bound by the ability of the calibrating student sample, and 4) the measures are linear so comparisons and growth in ability can be made.

To evaluate the reliability and validity of new items, every new scenario and item undergoes a science data review process. This process determines placement of new scenarios and items into the operational bank. New pilot item statistics are generated by administering a WASL "practice" test to approximately 1,200 students representing the diversity of the state. During the science data review process, test results from piloted items are used to study relationships between the items and the test takers. Certain criteria must be met before new scenarios and items are recommended for the operational bank (Willhoft & Yin, 2005). Criteria for an acceptable item include 50% or more responses were correct for items with multiple-choice response format (*p*-value = 0.50) and 50% or more of the points are earned on short and extended constructed response format (*p*-value = 1.0 for short answer, *p*-value = 2.0 for extended response). Point biserial (correlation) measures how students performed on the item compared to their performance on all the items. The theoretical range is from -1.0 to +1.0. A positive point biserial means test takers scoring higher on the exam were more likely to answer the item correctly. A negative point biserial means test takers scoring higher on the exam were more likely to answer the item incorrectly. Items are acceptable when point biserial values are positive and > 0.25.

Differential item functioning (DIF) methods are used to examine response patterns across ethnicity, gender and language groups. The expectation is that an individual's ability dominates the measure and situational variables such as gender, ethnicity, and linguistic group are minimal influences. DIF methods are used in item calibration steps. Ideally, there should be no performance difference between groups. New items that meet these criteria may be included in an operational exam. After administration of an operational exam, student responses on the pilot items are used to calibrate an item difficulty parameter (R. Beven, personal communication, December 18, 2006).

Based on Item Response Theory, an Item Characteristics Curve (ICC) is created for every item on the operational exam. An ICC is derived from a response function, a plot that represents the way the item elicits responses from test takers with varying levels of ability. Possible values range from 0.0 to 1.0, representing the probability that a person of (theta sub i) ability gives a correct answer to an item with (b sub j) difficulty. Test scores are not linear, so to transform test scores into measures that approximate linearity, values are converted to the log odds ratio and recast into the final ICC. This is the form commonly used in the one-parameter logistic Rausch Model. Before the model is accepted, iterative procedures are used to simultaneously adjust both person and item parameters to obtain the best fit to the Rausch Model (Willhoft & Yin, 2005). The fit of the model is judged in several ways including evaluation of unidimensionality of data and Analysis of Squared Residuals summed across persons and items.

An example ICC is illustrated in Appendix D. The y-axis represents the probability of a correct response ranging from 0.0 to 1.0. The x-axis represents the item difficulty value ranging from -4.0 to 4.0. An item difficulty value is assigned to each item by tracing the curve upward to the point where y = 0.5 (the probability of a correct answer = 0.5.) The greater the value, the more difficult the item. Once items are calibrated with an item difficulty value between -4 and +4, they are placed in a calibrated item bank.

IRT makes possible certain assumptions about student test responses; therefore, responses to each item are compared to assumed expected responses. Item fit statistics provide information on item response patterns. Construct validity is tested by examining these item fit statistics. A typical item response pattern is one that is relatively orderly. Response patterns with many outlier responses confound efforts to identify a core set of items that is a valid measure of a particular construct. Expected and desired responses are those where the probability of correct responses increase as ability level increases. This is called positive item discrimination. Negative item discrimination is observed when the probability of a correct response decreases as ability level increases. An example of positive and negative item discrimination appear in Appendices E and F. Items with negative discrimination are not desirable. When item responses do not fit the measurement model, the model is not compromised or discarded. Rather, better items are sought that produce consistent responses in ability levels across all gender, ethnic, and language groups. Each year, two subsets are constructed into Form A and Form B of the science WASL. Items are selected for scenario and stand-alone question types based on item difficulty values and assembled into the WASL using Big Steps software (R. Beven, personal communication, December 18, 2006).

Data Sources

The testing instrument will be the 2005 Science Washington Assessment of Student Learning (WASL). The Office of the Superintendent of Public Instruction will provide an SPSS file of 2005 8th grade science WASL scores for each test item. In addition, ethnicity and gender will be indicated on each student record.

The 2005 8th-grade science WASL had 45 total items: 39 items were presented in a scenario question type and 6 items were presented in a stand-alone question type. The items were either multiple-choice or short or extended constructed response format. The multiple-choice items were worth 2 points each. The short constructed response items were scored on a partial credit scale (0,1,2). The extended constructed response items required detailed diagrams and longer text responses and were graded on a partial credit scale (0,1,2,3,4). Interrater reliability was reported at 90%, using the percentage agreement method.

Instrument Reliability

Cronbach's Alpha is a measure of internal consistency used to help determine which items become operational and which items are rejected. When Cronbach's Alpha

> 0.85, it is assumed that only one theoretical construct is being measured (Schmitt, 1996). On the 2005 8th-grade science WASL, Cronbach's alpha = 0.89, indicating that the test is unidimensional.

Templates are used to map test items to the construct intended to be measured. Templates include the scenario name, a text description of the construct, the EALR and GLE for each item, and item type. The construct template used to guide the design of scenario "In the Doghouse" is included in Appendix G.

Preparation of Data before Analysis

This research required many hours of collaboration with staff in the Information Technology Department at Washington State's Office of the Superintendent of Public Instruction. A collaborative effort was required to create a database extract file that contained pertinent student data and valid student records. This section provides an overview of the process and decision points that shaped the research questions.

Washington State's assessment data are stored in a relational database. A relational database is a collection of data arranged in logical structures called tables. Table structures are designed to assist in quick retrieval of information that results in minimal processing response time. This is partly achieved by creating tables that hold related data (hence the name relational database). Within a table, data reside in placeholders called fields. For example, one table may hold student demographic data and have field names such as first_name, last_name, gender, and ethnicity. Another table may hold assessment results and include field names such as test_date, content_area, test_score, and proficiency_level.

Programmers use a computer language, such as SQL, to extract data from database tables. The extracted data is usually stored in a comma-delimited file that can be imported into another application, such as SPSS or Excel. Due to the large number of fields or variables in each table, it is more expeditious to write a SQL statement to extract all fields into a file rather than to type each individual variable name in a SQL statement. In other words, a SQL statement to extract a few fields may look like this: Select first_name, last_name, grade, teacher, gender, ethnicity, free/reduced lunch from table_name. In comparison, a SQL statement to extract all fields may look like this: Select * from table_name. Although this approach takes less effort for the programmer, it results in an arduous process for the researcher to filter through these data and determine pertinent fields. Often tables include over 100 fieldnames. The file extract used in this research was no exception as it contained 119 fields.

The next pre-analysis step was the development of a data dictionary to interpret often cryptic field names and values. Usually a data dictionary maps each variable name to an extended text definition. For example, the data dictionary entry for the disabilities code fieldname contains a key to interpret each of the 14 possible codes. If the disability code equals nine, this indicates that the student has autism. The creation of the data dictionary required multiple conversations with OSPI as well as the contractor, Pearson Consulting, who scored the 2005 WASL exams. Only after each party had a firm understanding of the meaning of each field in the extracted database file, could discussions ensue that would ultimately lead to the identification of fields needed for this analysis.

After the pertinent fields were identified, the next decision centered on which students to include in the analysis. Washington State education policies and procedures influenced these decisions. For example, one initial goal of this research was to gain an understanding of how Limited English Proficiency (LEP) students perform on different question types. However, Washington State policy gives students with limited English Proficiency (LEP) the choice to claim an exemption from the WASL exam. Consequently, very few LEP students elect to take the WASL exam and instead choose to wait until they reach a certain level of English proficiency and are no longer designated as LEP.

A further consideration was the inclusion of students with disabilities that elected to take the WASL. Washington State policy gives students with disabilities the choice to take the WASL exam with or without accommodations or to take the Washington Alternative Assessment (WAAS). Proficiency levels (cut scores) on the WASL for students with disabilities can be different from proficiency levels for students without disabilities. This decision is made by the district Special Education Director for each individual student. Approximately 14 different disability codes are used to designate a student's disability. Given the complexity of these factors, it was decided that students with disabilities warrant a separate study and therefore were not included in the data analysis.

After criteria for inclusion was defined, student records were coded as either "Y" (include) or "N" (do not include). This code also eliminated student records that were invalid due to missing values in the gender or ethnicity field, incomplete exams, or cheating.

Next, two new variables were created that consisted of the total raw score points on scenario question types and total raw score points on stand-alone question types. The creation of this variable required multiple steps in SPSS. First, each of 44 items received a designation based on whether the item appeared within a scenario or stand-alone question type. A dichotomous variable named q_type was created (1 = scenario, 2 = stand-alone). Second, each item received a designation based on item response format. A dichotomous variable named i_format was created (1 = multiple-choice, 2 = constructed response). Question type and item response format were determined by studying each item in the 2005 Science WASL test booklet and populating the SPSS spreadsheet with the correct value. Third, variables named scenraw and stdalraw were created to hold the total number of raw points each student received on scenario question types and stand-alone question types. The number of points awarded to each student for each item was calculated based on a code key obtained from Pearson. The code key specified the following logic rules:

1) If the item response format is multiple-choice and the student response equals A, B, C, or D, then the student receives 2 points.

2) If the item response format is multiple-choice and the student response is not equal to A, B, C, or D then the student receives 0 points.

3) If the item response format is short or extended constructed response and the student response is alphabetical (A-Z) then the student receives 0 points.

4) If the item response format is short or extended constructed response and the student response is not alphabetical (A-Z) then the student receives the number of points indicated in the field.

Using the variable i_format and the preceding logic, the total number of raw points for each item for each student record was calculated and entered in variables scenraw and stdalraw. These two variables became the dependent variables in this analysis.

Lastly, raw scores are not interpretable without first converting them to some relative measure; therefore, raw scores were converted to z-scores by using a linear transformation. The resulting z-scores retained the same magnitude of differences as the original raw scores (Anatasi & Urbina, 1997). *Z*-scores on scenario and stand-alone question types were further converted to normalized scores, or *T*-scores. *T*-scores make the comparison of scores on different tests valid. This procedure, first proposed by W.A. McCall (1922), is a convenient scale to use and recommended when sample sizes are large and representative of the population (Anatasi & Urbina, 1999), two conditions satisfied in this research.

The preceding discussion illustrates a few examples of how policy and available data shaped this research. The duration of time that elapsed from the development of the initial research questions to the final questions with a complete supporting data file was approximately 24 months.

Data Analysis

Students were grouped and coded based on ethnicity. Ethnicity was coded as a continuous variable with '1' = American Indian/Alaska Native, '2' = Asian, '3' = Black African American, '4' = Hispanic, '5' = White, '6' = Hawaiian Pacific Islander, and '7' = Multiracial. Gender was coded as a dichotomous variable with '1' = female and '2' = male.

The first question asked, "How do ethnic and gender groups perform on the 8th-grade science WASL?" A 2 (gender) X 7 Multivariate Analysis of Variance (MANOVA) was conducted to determine the effects of gender and ethnicity on 2005 eighth grade science WASL achievement scores. The combined dependent variable for this analysis consisted of student responses from scenario and stand-alone question types.

The second and third question asked, "How do ethnic and gender groups perform on scenario question types?" and "How do ethnic and gender groups perform on stand alone question types?" Two scores were calculated for each student: 1) a score of awarded points on items within scenario question types only and 2) a score of awarded points on items in stand-alone question types only. Each question type used different score scales because of the differences in the number of items and the score ranges. Following Allen & Yen (1979), both sets of raw scores were converted to a common scale for comparison purposes. First, raw scores were converted to Z-scores, then to *T*-scores with a mean of 50 and standard deviation of 10. Student scores on both scenario and stand-alone question types are reported on a common *T*-score scale for easier interpretability.

Total raw points on stand-alone question type and total raw points on scenario question type are the dependent variables and ethnicity and gender are the independent variables. A two-way MANOVA was used to study patterns of ethnicity and gender differences across these two question types. For each item, MANOVA tested for differences between ethnicities and genders across both scenario and stand-alone question types. If a significant interaction effect was obtained, MANOVA was followed with separate one-way ANOVAs to determine where the significant group differences existed across and within question type.

Effect Size Analysis

Because the dataset was very large (N=77,679), differences in achievement scores are more likely to be statistically significant. Large data sets reduce the standard errors to such a large magnitude that small trivial group differences frequently are statistically significant (Fields, 2000). Therefore, two measures of effect size, eta square (η^2) and Cohen's *d*, were reported when discussing the results of the analysis of variance. First, eta square values denote the proportion of variance in the dependent variable that can be explained by group differences or the strength of the relationship between the independent and dependent variables (Kennedy, 1970). Second, Cohen's *d* is a measure of the magnitude of differences between means in terms of standard deviation units (Cohen, 1988). Cohen's *d* was calculated by subtracting the two means of interest and dividing by their pooled standard deviation.

CHAPTER 4

RESULTS

This chapter presents the results of the study. The first section presents students' patterns of performance on the 2005 eighth-grade science WASL. Results are reported for overall test scores and then disaggregated by question type across ethnicity, gender, and gender within ethnicity. Two measures of effect size, eta square (η^2) and Cohen's *d*, are reported when discussing significant main effects and significant interactions. The second section provides the results in terms of analysis of variance.

Students' Patterns of Performance

Overall Test Performance by Ethnicity

The first research question was designed to identify how ethnic and gender groups perform on the 2005 8th-grade science WASL. Analyses indicated that overall, approximately 38% of the students scored at proficient and above proficient levels. The remaining 62% of the students scored below proficient levels and thus, did not meet the state standards. Proficiency levels are divided into four levels: level 1 denotes below basic, level 2 denotes basic, level 3 denotes proficient, and level 4 denotes above proficient.

There were large differences in the percentage of students proficient and not proficient within each ethnicity (Figure 3). Approximately 40% of the students were proficient in the top three ranking ethnic subgroups (Multiracial (MR), White, Asian).

This figure dropped substantially in the remaining subgroups where approximately 20% and less of the students were proficient. A majority of the students in the lower performing subgroups (American Indian/Alaska Native, Black/African American, Hispanic, and Hawaiian/Pacific Islander) were in level 1, the lowest proficiency level (Table 2).

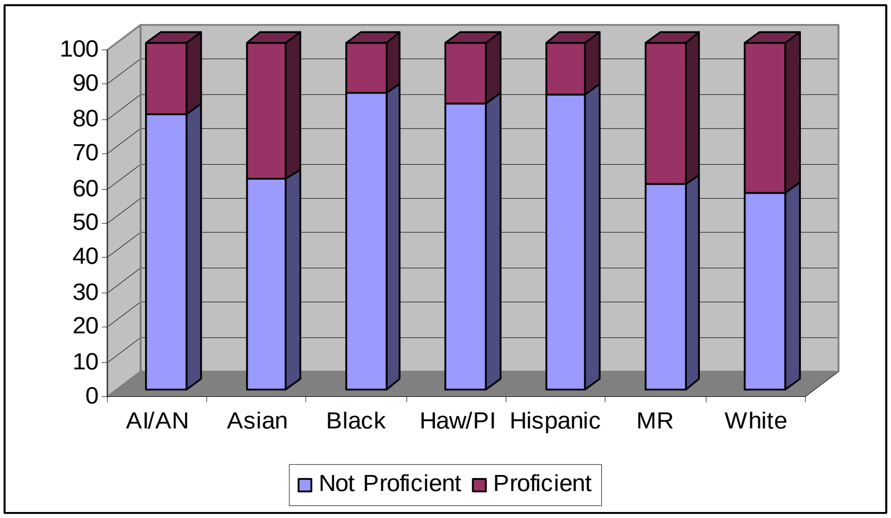


Figure 3. Percentage of Students Proficient and not Proficient by Ethnicity.

Table 2. Percentage of Students in each Proficiency Level by Ethnicity.

	Proficiency Level			
Ethnicity	*Level 1 (Lowest)*	*Level 2 (Not Proficient)*	*Level 3 (Proficient)*	*Level 4 (Highest)*
Hispanic	52.2%	33.0%	13.9%	0.9%
Black	51.9%	33.4%	14.1%	0.6%
Hawaiian/PI	46.4%	36.2%	17.4%	0.0%
AI/AN	42.6%	36.8%	18.9%	1.7%
Asian	25.1%	35.4%	33.7%	7.2%
Multiracial	24.2%	34.8%	33.7%	7.2%
White	20.4%	36.4%	37.1%	6.1%

Overall Test Performance by Gender

Eighth grade girls and boys had similar scores on the 2005 science WASL. Thirty-seven percent of girls met state standards and 38% of boys met state science standards. Distributions within each proficiency level were also similar (Table 3).

Table 3. Percentage of Students in each Proficiency Level by Gender.

	Proficiency Level			
Gender	*Level 1* *(Lowest)*	*Level 2* *(Not Proficient)*	*Level 3* *(Proficient)*	*Level 4* *(Highest)*
Female	28.3%	34.7%	31.8%	5.2%
Male	25.2%	36.9%	33.0%	4.9%

Performance within Scenario Question Type

In order to gain further understanding of students' performance, this study examined how ethnic and gender groups performed on the two different types of questions (scenario and stand-alone) presented on the WASL. The raw scores generated by the scenario and stand-alone question types used different score scales. Scenario question types included four to six separate items and scores ranged from 0-11 points while stand-alone question types included one item and scores ranged from 0-2 points. Due to the differences in the number of items and the score ranges, both sets of raw scores were converted to a common scale or standardized for comparison purposes. All raw scores were converted to Z-scores, then to *T*-scores with a mean of 50 and standard deviation of 10 (Allen & Yen, 1979). Student scores on both scenario and stand-alone question types are reported using *T*-scores for easier interpretability. When evaluating educational significance, it is important to note that each *T*-score point is the equivalent of three to seven scaled score points. This is due to IRT-based equating procedures used to convert raw scores to scaled scores. In this procedure, a value of theta is calculated based on the relationship between raw score and ability. Each raw score is then multiplied by the theta value to determine the number of scaled score points corresponding to each raw score. OSPI establishes proficiency levels based on the number of scaled score points. It is not possible to report the percentage of students within each proficiency level by question type because established cut-scores apply to total scaled score points only.

Performance by Ethnicity within Scenario Question Type

In order to gain further understanding of students' performance, this study examined how ethnic and gender groups performed on the scenario question type. This question format presents five to six items (multiple-choice, short or extended constructed response format) within a story context that provides an authentic, real-life application. Table 3 lists the mean *T*-score and standard deviation for each ethnicity. In general, mean scores were low. White students had the highest mean score (51.60), followed by Multiracial students (51.00), and Asian students (50.69). American Indian/Alaska Native and Hawaiian/Pacific Islander students had similar mean scores of 45.67 and 45.32, respectively. Black African American and Hispanic students had the lowest mean scores of 43.64 and 43.46, respectively.

Table 4. Mean T-Score and Standard Deviation by Ethnicity for Scenario Question Types.

Ethnicity	*N*	Scenario *Mean*	*SD*
American Indian/Alaska Native	2189	45.67	9.73
Asian	6043	50.69	9.93
Black African American	4324	43.64	9.44
Hispanic	8709	43.46	9.66
White	56,094	51.60	9.44
Hawaiian/ Pacific Islander	69	45.32	9.45
Multiracial	264	51.00	9.85

The effect of scenario question types was most striking for Hispanic and Black African American students. An eight-point spread separated the highest mean scenario score (51.60), achieved by White students, and the two lowest mean scores (43.46 & 43.64) achieved by Hispanic and Black African American students. This result was investigated using eta square values and Cohen's *d* values. Eta square denotes the proportion of variance in the dependent variable explained by group differences or the strength of the relationship between the independent and dependent variables (Kennedy, 1970). The partial eta squared value for the independent variable ethnicity and the dependent variable mean scores on scenario question types was .095. This indicates that almost 10% of the variability in mean scores on scenario question types was explained by ethnicity.

Cohen’s *d* is a measure of the magnitude of differences between means in terms of standard deviation units (Cohen, 1988). Cohen’s *d* was calculated by subtracting the two means of interest and dividing by their pooled standard deviation. Cohen’s *d* values ranging between 0.2 and 0.4 are considered a small effect, values ranging between 0.5 and 0.7 are considered medium, and values 0.8 and above are considered a large effect.

In this study, large effect sizes (Cohen’s $d > .80$) were found when comparing mean scenario scores for 1) White and Hispanic students and 2) White and Black African American students (Table 5). It may be helpful to think of Cohen’s *d* as the mean difference between two variables expressed in standard deviation units. Thus, White students on average scored at or above .80 standard deviation unit higher than either Hispanic or Black African American students. This indicates a large magnitude of difference in performance on scenario question types between 1) White and Hispanic students and 2) White and Black African American students.

Similarly, substantial effect sizes were found when comparing mean scenario scores for Multiracial and Hispanic students (Cohen’s $d = .76$) and Multiracial and Black African American students (Cohen’s $d = .77$) (Table 5). Multiracial students’ mean score (51.00) was higher than Hispanic and Black African American students’ mean score (43.46 and 43.64, respectively). Finally, substantial effect sizes were found when comparing mean scenario scores for Asian and Hispanic students (Cohen’s $d = .74$) and Asian and Black African American students (Cohen’s $d = .73$). Again, Asian students’ mean score (50.69) was higher than Hispanic and Black African American students’ mean score (43.46 and 43.64, respectively). This trend suggests that the effect of scenario question types on achievement scores of Hispanic and Black students is notable.

Effect sizes of moderate magnitude were found between the following subgroups: 1) Multiracial and Hawaiian/Pacific Islander, 2) Multiracial and American Indian/Alaska Native, 3) Asian and Hawaiian/Pacific Islander, 4) Asian and American Indian/Alaska Native, 5) White and American Indian/Alaska Native subgroup. While these effect sizes were not large, they still imply a non-trivial relationship between performance on scenario question types and ethnicity.

Table 5. Effect Size by Ethnicity for Scenario Question Type.

Ethnicity	*Asian*	*Black*	*Hispanic*	*White*	*HPI*	*MR*
AI/NA	0.510	0.212	0.229	0.619	0.037	0.543
Asian		0.727	0.738	0.095	0.553	0.031
Black			0.020	0.843	0.178	0.762
Hispanic				0.854	0.196	0.773
White					0.665	0.064
HPI						0.587

Table 6 presents the results of dividing students' scores into two groups: students who were awarded less than 50% of the total points possible and students who were awarded 50% or more of the points possible on the scenario question type. The table disaggregates these data by ethnicity. White, Multiracial, and Asian subgroups performed similarly with approximately 20% of the students earning 50% or more of the total points. American Indian/Alaska Native and Hawaiian/Pacific Islander subgroups had similar performance with approximately 8% of the students earning 50% or more of the total points. Note in the Black and Hispanic subgroup, fewer than 5% of the students were awarded 50% or more of the total points possible.

Table 6. Percentage of Students in each Category by Ethnicity for Scenario Question Types.

	Scenario Question Type	
Ethnicity	< 50% Correct	> or =50% Correct
American Indian	91.4%	8.6%
Asian	80.6%	19.4%
Black	95.4%	4.6%
Hispanic	95.1%	4.9%
White	78.8%	21.2%
Hawaiian/Pacific Islander	92.8%	7.2%
Multiracial	78.4%	21.6%

The box-and-whiskers plot in Figure 4 shows that roughly 75% of the Hispanic and Black African American students scored less than the mean score of the White and Multiracial students on scenario question types.

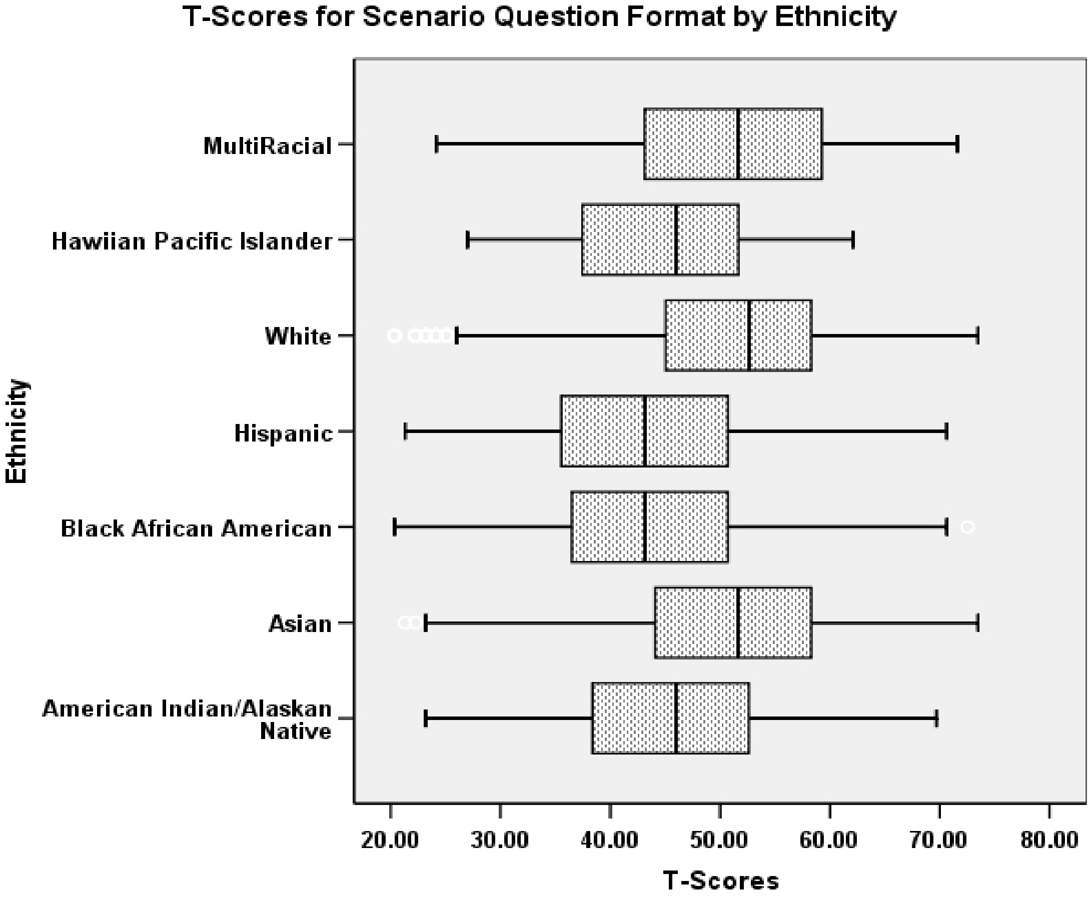


Figure 4. Box-and-whiskers Plot of T-scores by Ethnicity for Scenario Question Types.

Performance by Gender within Scenario Question Type

When the data were disaggregated by gender, the results indicated that although not significantly different, girls' mean score was higher than boys' mean score within scenario question types (50.40 and 49.62, respectively). Girls' scores were more tightly clustered around the mean compared to boys' scores (9.74 and 10.23 standard deviation units). The gender by ethnicity interaction effect was significant for American Indian/Alaskan Native, Asian, Black African American and White subgroups, indicating that, within these subgroups, boys' and girls' scenario mean scores are statistically significantly different. Table 7 shows the difference between girls' and boys' mean *T*-score by ethnicity for the scenario question type. Effect size, Cohen's *d*, is also reported in the table. Note that all values of Cohen's *d* are less than 0.20. Partial eta square value equals 0.00. This indicates that, from a practical standpoint, the magnitude of difference in mean scores between genders within ethnicity is negligible.

Gay (2000) urged researchers to study gender differences within each ethnicity due to cultural differences; therefore, it is notable that girls outperformed boys in all ethnicities except for Hawaiian/Pacific Islander. The largest point difference was between Black African American girls and boys. Black/African American girls performed nearly 1.75 points higher than boys on scenario question types. In the American Indian/Alaska Native subgroup, girls scored almost 1.25 points higher than boys. In the Asian and Multiracial subgroups, girls scored slightly more than 0.50 point higher than boys. In the White subgroup, girls scored slightly more than 0.75 points higher than boys. The smallest point difference (0.18) was between Hispanic boys and girls.

Table 7. Difference on Girls' and Boys' Mean T-score by Ethnicity for Scenario Question Types.

Ethnicity	*Mean Score Difference*	*Cohen's d*
American Indian/Alaska Native	1.20	0.12
Asian	0.61	-0.06
Black African American	1.70	0.18
Hispanic	0.18	0.02
White	0.85	0.09
Hawaiian/ Pacific Islander	-1.20	0.13
Multiracial	0.64	0.06

Performance within Stand-Alone Question Type

Performance by Ethnicity within Stand-Alone Question Type

The analysis also examined ethnicity and gender differences in performance within the stand-alone question types. This question format presents one item (multiple-choice or short constructed response format) that is not introduced within a story context. Table 8 contains mean *T*-scores and standard deviations by ethnicity. In general, mean scores were low. White students had the highest mean score (51.37), followed by Multiracial students (51.28), and Asian students (49.93). American Indian/Alaska Native mean score was 46.90. Hawaiian/Pacific Islander, Black African American and Hispanic students had similar mean scores of 44.50, 44.31 and 44.83, respectively.

Interestingly, the top three ethnic subgroups on scenario question types are identical to the top three ethnic subgroups for stand-alone question types (i.e., White, Multiracial, and Asian subgroups). For the first two subgroups, White and Multiracial, there is little difference between the mean score on scenario question types and the mean score on stand-alone question types. The largest difference between these question types was realized by the lowest scoring subgroups. For example, in the Hispanic subgroup, the mean number of score points achieved on stand-alone question types was 1.37 points higher than the mean number of score points achieved on scenario question types. Similarly, in the American Indian/Alaska Native subgroup, the mean number of score points achieved on stand-alone question types was 1.23 points higher than the mean number of score points achieved on scenario question types. Black/African American students earned 0.67 points more on stand-alone question types compared to scenario question types. Asian students scored 0.76 points lower and Hawaiian Pacific Islander students scored 0.82 points lower on stand-alone versus scenario question types.

Table 8. Mean T-Score, Standard Deviation and Difference in Mean Stand-alone vs. Scenario Scores.

Ethnicity	*Stand-Alone Mean*	*SD*	*Δ Score Points*
American Indian/Alaska Native	46.90	9.97	+1.23
Asian	49.93	9.91	-0.76
Black African American	44.31	10.18	+0.67
Hispanic	44.83	10.08	+1.37
White	51.37	9.53	-0.23
Hawaiian/ Pacific Islander	44.50	9.70	-0.82
Multiracial	51.28	10.40	+0.28

The box-and-whiskers plot in Figure 5 shows a positive shift toward the right in the distribution of scores for Hispanic, Black African American, and American Indian/Alaska Native students (Figure 5). This shift indicates that a larger portion of lower performing students approached the mean score of White students. The simultaneous decrease in the mean score of White students and the increase in the mean score of lower performing subgroups lessened the achievement gap.

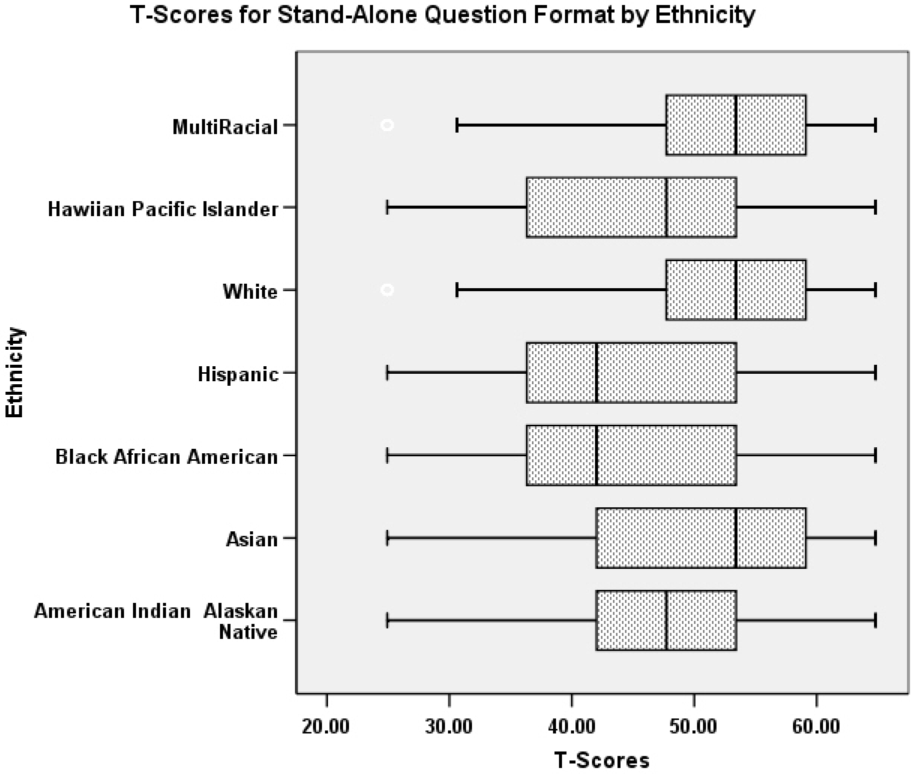


Figure 5. Box-and-whiskers Plot of T-scores by Ethnicity for Stand-alone Question Types.

The advantage the higher performing subgroups (White, Asian, Multiracial) had over the Black/African American and Hispanic subgroups was not as profound on stand-alone question types compared to scenario question types. Large effect sizes found on scenario question types, particularly between Hispanic and Black African American students, were smaller on stand-alone question types. The point difference between the highest mean stand-alone score (51.37), achieved by White students, and the two lowest mean scores (44.83 and 44.31) achieved by Hispanic and Black African American students, was also smaller than the point difference between these groups on scenario question types. Moderate effect sizes (Cohen's *d* < .70) were found when comparing mean scenario scores for White and Hispanic students and White and Black African American students (Table 9). Thus, White students on average scored approximately .70 standard deviation units higher than either Hispanic or Black African American students. Similarly, smaller effect sizes were found when comparing mean stand-alone scores to mean scenario scores for Multiracial and Hispanic students (Cohen's *d* = .62) and for Multiracial and Black African American students (Cohen's *d* = .67). Multiracial students' mean score (51.28) was higher than Hispanic and Black African American students' mean score (44.31 and 44.83, respectively). Finally, substantial effect sizes between Asian and Hispanic or Black African American students on scenario scores were reduced to very moderate when comparing mean stand-alone scores for Asian and Hispanic students (Cohen's *d* = .51) and for Asian and Black African American students (Cohen's *d* = .57). Asian students' mean stand-alone score (49.93) was higher than Hispanic and Black African American students' mean stand-alone scores (44.83 and 44.31, respectively) but the difference in stand-alone mean scores was smaller than the difference in scenario mean scores.

These data indicate that there was less deviation in mean scores between ethnic subgroups on stand-alone question types compared to scenario question types. In other words, the magnitude of difference in mean scores between ethnic subgroups was smaller on stand-alone question types as compared to scenario question types. The partial eta squared value for stand-alone question types was small (6.5%), indicating that roughly 6.5% of the variability in mean scores on stand-alone question types was explained by ethnicity.

Table 9. Effect Size by Ethnicity for Stand-Alone Question Type.

Ethnicity	Asian	Black	Hispanic	White	HPI	MR
AI/NA	0.287	0.282	0.227	0.409	0.290	0.394
Asian		0.567	0.513	0.117	0.581	0.113
Black			0.056	0.694	0.000	0.669
Hispanic				0.640	0.058	0.616
White					0.712	0.000
HPI						0.686

Table 10 compares the results of students who were awarded less than 50% of the total points possible and students awarded 50% or more of the points possible on stand-alone question types. The table further disaggregates these data by ethnicity. Recall that in the Black and Hispanic subgroup, fewer than 5% of the students were awarded 50% or more points on scenario question types. On stand-alone question types, this percentage makes a jump to just under 15% of the students that were awarded 50% or more points. This trend appeared in all ethnicities. In other words, regardless of ethnicity, more students received a greater percentage of points on stand-alone question types compared to scenario question types.

Table 10. Percentage of Students by Ethnicity Achieving Above and Below 50% Correct on Stand-alone Question Types.

	Stand-Alone Question Type	
Ethnicity	< 50% Correct	> or =50% Correct
American Indian	80.0%	20.0%
Asian	69.1%	30.9%
Black	85.9%	14.1%
Hispanic	85.5%	14.5%
White	64.8%	35.2%
Hawaiian/Pacific Islander	91.3%	8.7%
Multiracial	65.9%	34.1%

Performance by Question Type within Ethnicity

Performance between question types within each ethnic subgroup was examined by calculating the difference (in standard deviation units) between scenario and stand-alone T-score points. The magnitude of difference in performance on each question types within ethnicity was determined by first calculating the difference between scenario T-score points and stand-alone T-score points for each student. Next, the mean of the differences was calculated for each ethnic subgroup. Finally, this mean was divided by the standard deviation of each ethnic subgroup to arrive at Cohen's *d*.

Table 11. Binomial Effect Size Display (BESD) by Ethnicity.

Ethnicity	Cohen's *d*	*r*	N	# of Students Affected
American Indian/ Alaskan Native	0.16	.08	2,189	175
Asian	0.10	.05	6,043	302
Black	0.09	.05	4,324	216
Hispanic	0.17	.09	8,709	784
White	0.03	.01	56,094	1,122
Hawaiian/Pacific Islander	0.10	.05	69	3
Multiracial	0.04	.02	264	5

Binomial effect size display (BESD) is a useful method to examine this difference by providing a context and a practical interpretation (Rosenthal & Rubin, 1982). Table 11 lists the BESD for each ethnic subgroup. The effect size difference in the American Indian/Alaskan Native and Hispanic subgroups can be interpreted as approximately 8%

more American Indian/Alaska Native (175) and 9% more Hispanic (784) students scored higher on stand-alone compared to scenario question types. Approximately 5% more Black (216) students and approximately 2% more Multiracial (5) students scored higher on stand-alone compared to scenario question types. Although these numbers are small, they still can be meaningful in the context of student achievement on high-stakes tests.

On the other hand, Asian, Hawaiian Pacific Islander, and White and students favored scenario question types. Using BESD, this can be interpreted as approximately 5% Asian (302) students, 5% Hawaiian Pacific Islander (3), and 1% more White (1,122) students scored higher scores on scenario compared to stand-alone question types.

Performance by Gender within Stand-Alone Question Types

Stand-alone question types yielded a similar gender pattern to scenario question types. Although not significantly different, girls' mean score was higher than boys' mean score (50.27 and 49.74, respectively). Once again, girls' scores were more tightly clustered around the mean compared to boys' scores (9.82 and 10.17 standard deviation units). The gender by ethnicity interaction effect was significant for American Indian/Alaskan Native, Black African American and White subgroups, indicating that boys' and girls' mean score on stand-alone question types were significantly different. Note that all values of Cohen's *d* are 0.16 or less, indicating from a practical standpoint, the difference in mean scores between genders within ethnicity is negligible. Partial eta square value equals 0.00 indicating no relationship between gender and performance on stand-alone question types.

In general, girls scored higher on stand-alone question types in all ethnicities except for Hawaiian Pacific Islander. The largest score point differences were found between genders in the American Indian/Alaska Native, Black African American, and Hawaiian/Pacific Islander subgroups (Table 12). In the Black African American, White, and Multiracial subgroups, the score point advantage that girls had was not as large on stand-alone versus scenario question types. In contrast, in the Hawaiian Pacific Islander subgroup, the score point advantage that girls had was larger on stand-alone versus scenario question types. Boys and girls in the remaining subgroups had approximately the same mean score differences on stand-alone versus scenario question types.

Table 12. Mean T-Score Gender Differences by Ethnicity for Stand-Alone Question Type.

Ethnicity	Mean Score Difference	Cohen's *d*
American Indian/Alaska Native	1.23	0.12
Asian	0.61	0.00
Black African American	1.24	0.12
Hispanic	0.22	0.02
White	0.60	0.07
Hawaiian/ Pacific Islander	-1.58	0.16
Multiracial	0.14	0.02

Performance with Constructed Response Items Removed

Performance by Ethnicity

There were no students designated as English Language Learners (ELL) in the data set; however, students have varying levels of reading and writing ability. To mediate the potentially confounding effects of the extended constructed response items within scenario question types, additional analyses were conducted after removing short and extended constructed response items from both question types. The analysis examined ethnicity and gender differences in performance within scenario and stand-alone question types on multiple-choice items only.

The same patterns of performance emerged with and without the constructed response items included in the analysis (Table 13). Once again, the top three scoring ethnic subgroups on scenario question types are identical to the top three scoring ethnic subgroups for stand-alone question types (i.e., White, Multiracial, and Asian subgroups). There were slight differences in mean scores on scenario question types compared to mean scores on stand-alone question types for the White and Multiracial subgroups. The largest increase in mean score on stand-alone question types (multiple-choice items only) compared to the mean score on scenario question types (multiple-choice items only) was realized by the lowest scoring subgroups. For example, in the Hispanic subgroup, the mean number of points achieved on stand-alone question types was 1.86 points higher than the mean number of points achieved on scenario question types. In the American Indian/Alaska Native subgroup, the mean number of points achieved on stand-alone question types was 1.19 points higher than the mean number of score points achieved on scenario question types. Black/African American students earned 0.81 points more on stand-alone question types compared to scenario question types. Similar to the previous analysis, Asian students scored 0.60 points lower and Hawaiian Pacific Islander students scored 1.06 points lower on stand-alone versus scenario question types.

Although the difference between the mean score on scenario question types and the mean score on stand-alone question types for the White and Multiracial subgroups is small, mean scores on stand-alone questions did decrease. Mean scores on stand-alone questions also decreased for the Asian and Hawaiian Pacific Islander subgroups. In contrast, mean scores on stand-alone questions for the remaining subgroups increased. The simultaneous decrease in mean scores on stand-alone question types for higher performing subgroups and increase in mean scores on stand-alone question types for lower performing subgroups had the effect of lessening the achievement gap.

Table 13. Means and Standard Deviations on Multiple-choice Items only by Ethnicity.

	Mean *T*-Score M/C Scenario		Mean *T*-Score M/C Stand-Alone	
Ethnicity	*M*	*SD*	*M*	*SD*
American Indian/Alaskan Native	46.19	10.47	47.38	10.36
Asian	50.08	10.00	49.48	9.99
Black African American	43.90	10.22	44.71	10.50
Hispanic	43.14	10.39	45.00	10.25
White	51.68	9.19	51.35	9.49
Hawaiian Pacific Islander	45.22	10.14	44.16	9.85
Multiracial	51.36	9.56	51.00	10.07

Partial eta squared values between ethnicity and achievement scores by question type are low and nearly identical for both analyses. The ethnicity factor accounted for 10% of the overall variance in achievement scores on scenario question types with and without constructed response items. The ethnicity factor accounted for 6% of the overall variance in achievement scores on stand-alone question types without constructed response items and 7% of the overall variance in achievement scores on stand-alone question types with constructed response items. This means that, in general, the strength of the relationship between achievement scores within ethnic subgroups and question type is weak. Further, the relationship does not change whether constructed response items are included or removed from the analysis.

Performance by Gender

Both girls' and boys' mean scores decreased when constructed response items were removed; however, girls' mean scores decreased more than boys' mean scores. Girls' mean scores decreased between 3.0 and 3.5 points while boys' mean scores decreased between 1.5 and 2 points. Boys' mean score was higher than girls' mean score within scenario question types (48.00 and 46.71, respectively) and stand-alone question types (47.89 and 47.26, respectively). Similar to the previous analysis that included both constructed response and multiple-choice items, girls' scores were more tightly clustered around the mean compared to boys' scores on both question types (Table 14). Partial eta square values between gender and achievement scores by question type are identical for both analyses (0.00), indicating that despite score differences between genders, these differences were not of a very large statistical magnitude.

Table 14. Means and Standard Deviations on Multiple-choice Items only by Gender.

	Mean *T*-Score M/C Scenario		Mean *T*-Score M/C Stand-Alone	
Gender	*M*	*SD*	*M*	*SD*
Girls	46.71	9.62	47.26	9.93
Boys	48.00	10.32	47.89	10.07

Results Reported in Terms of Analysis of Variance

The Pearson correlation coefficient was used to determine if a correlation existed between the two dependent variables, namely scenario and stand-alone item responses. The correlation between the scenario and stand-alone item responses for all students was .70, indicating that it was appropriate to combine the dependent variables for the MANOVA analysis. MANOVA uses two distinct steps to test for group differences. First, a preliminary step determines if group differences (ethnicity and gender) exist for the combined dependent variable (main effects). If there are significant main effects, the second step is to use follow up univariate ANOVAs to test the individual effects of each independent variable (ethnicity, gender) on the dependent variable (scenario points and stand-alone points).

A 2 (gender) X 7 (ethnicity: American Indian/Alaskan Native, Asian, Black African American, Hispanic, White, Hawaiian Pacific Islander, Multiracial) Multivariate Analysis of Variance (MANOVA) revealed no significant main effects for gender (Pillai's trace = .000, $F_{(1, 77678)} = 1.23$, $p = .293$, Partial Eta Square = .000) when the dependent variable was comprised of the combined scenario and stand-alone question types (Table 15). This means that girls' and boys' mean achievement scores on the overall test did not vary significantly. However, there were significant main effects for ethnicity (Pillai's trace = .099, $F_{(6,\ 77678)}$ = 675.21, p = .000, Partial Eta Square = .050). This means that mean achievement scores between ethnic subgroups varied significantly. The gender by ethnicity interaction effect was found to be significant (Pillai's trace = .000, $F_{(6,\ 77678)}$ = 2.48, p = .003, Partial Eta Square = .000). This means that within ethnicity, girls' and boys' mean achievement scores on the combined scores varied significantly.

Results from follow-up univariate ANOVAs indicated significant effects for ethnicity in both scenario ($F_{(6,\ 77678)}$ = 1365.17, p = .000, Partial Eta Square = .095) and stand-alone question types ($F_{(6,\ 77678)}$ = 897.22, p = .000, Partial Eta Square = .065). In addition, simple effects ANOVAs indicated significant gender by ethnicity interaction effects for both scenario ($F_{(13,\ 77678)}$ = 3.62, p = .001, Partial Eta Square = .000) and stand-alone question types ($F_{(13,\ 77678)}$ = 2.85, p = .009, Partial Eta Square = .000) (Table 15).

One of the assumptions that needs to be met for MANOVA to yield valid results is homogeneity of variances of each group (gender, ethnicity). However, the homogeneity assumption is not met because of the large number of White students (N=56,094) compared to the small number of Hawaiian Pacific Islander (N=69) and Multiracial students (N=264). The Pillai-Bartlett's trace is the most appropriate statistic to use when the sample sizes are unequal (Tabachnick & Fidell, 2007); therefore, Pillai-

Bartlett's trace F-statistic was used to detect group differences on the combined dependent variable. Multivariate normality was evaluated using Box's test. The Box test is very sensitive to large sample sizes and may find significant heterogeneity that will not impact the MANOVA analysis. To address the sensitive nature of this test, an exploratory analysis involving descriptive statistics and histograms was conducted for gender and ethnicity for both stand-alone and scenario question types. Skewness and kurtosis values across both question types for ethnicity and gender did not exceed 2.00 indicating that distributions were normal (Howell, 2002).

Table 15. Analysis of Variance for Question Type.

Source	*df*	*F*	*η*	*p*
Gender (G)				
Scenario	1	2.25	.000	.133
Stand-Alone	1	.481	.000	.488
Ethnicity (E)				
Scenario	6	1365.17	.095	.000
Stand-Alone	6	897.22	.065	.000
G X E		.		
Scenario	6	3.62	.000	.001
Stand-Alone	6	2.85	.000	.009

Summary

These data suggested that the 2005 eighth-grade science WASL was gender neutral. Girls' and boys' mean scores and standard deviations were very similar for both question types. Gender differences within ethnicity were also negligible. On the other hand, significant achievement differences between ethnic subgroups persisted for both question types.

The score advantage the higher performing subgroups (White, Asian, Multiracial) had over the lower performing subgroups (American Indian/Alaskan Native, Black/African American, Hispanic) was not as profound on stand-alone question types compared to scenario question types. Large standard deviation units between mean scenario scores of different ethnic subgroups became very moderate standard deviation units between mean stand-alone scores. A greater percentage of students in underrepresented groups received a greater percentage of score points on stand-alone question types compared to scenario question types. This trend suggests that stand-alone question types had a leveling effect across ethnicities even when the constructed response items were removed from the analyses. The performance patterns were consistent across multiple-choice and constructed response formats.

Scenario question types appeared to register differences between ethnic groups to a greater degree than stand-alone question types. These differences may be attributable to individual differences in cognition (Marshall, 1995), characteristics of test items themselves (Shaftel et al., 2006) and opportunities to learn (Darling-Hammond, 2000).

Indeed, performance on tests within the mathematics and science domain, has been found to be highly sensitive to differences in classroom instruction (D’Agostino, 2007).

CHAPTER 5

DISCUSSION

This chapter provides an overview of the purpose and significance of this study as well as a summary of major findings from the data analyses. This chapter also presents a discussion of the implications of the findings regarding performance differences within gender and ethnic subgroups and considers factors that may have contributed to differential performance. The chapter concludes with suggestions for educators and policy makers and provides suggestions for future research.

Purpose of the Study

Significant differences in achievement among ethnic groups persist on the eighth-grade science Washington Assessment of Student Learning (WASL). The WASL measures academic performance in science using two types of questions: scenario question types and stand-alone question types. Scenario question types consist of four to six separate items presented in an authentic performance-like context. These items have a multiple-choice format or short-constructed or extended-constructed response format. Stand-alone questions are not linked to a scenario and have either a multiple-choice or short constructed response.

Three kinds of scenarios are presented in the WASL and are modeled after the Systems, Inquiry, and Application Strands of the Washington State science standards. The Systems Strand consists of content knowledge about the properties, structure, and changes in physical, earth, space, and living systems. The Inquiry Strand describes components of problem solving needed to investigate the concepts and processes of nature. In other words, students are expected to engage in the process of inquiry to discover new knowledge. Lastly, the Application Strand, recently renamed the Design Strand, focuses on the cognitive skills needed to apply knowledge and design solutions to human problems dealing with science, technology, and societal issues. This holistic approach to science education provides teachers and students with a unified way of thinking.

Each strand includes both Essential Academic Learning Requirements (EALRs) and Grade Level Expectations (GLEs) in the science domain. The EALRs and GLEs prescribe what all students should know at each of these grade levels and reflect the model of cognition from which assessments are designed and created. These learning requirements and expectations in each strand reflect a developmental model of cognition depicting increasingly sophisticated stages of understanding as students move through elementary, middle, and finally high school. Within the GLEs, sections entitled "evidence of learning" describe students' advancement in domain knowledge and strategic processing as they move from a level of acclimation to one of proficiency.

Previous research indicates that the presentation of target items connected to an authentic context such as scenario question types, may increase science achievement scores especially in underrepresented groups and help to close the achievement gap (Stecher et al., 2000). Thus, the present study focused on identifying significant differences in performance between gender and ethnic subgroups on scenario and stand-alone question types on the 2005 eighth-grade science WASL. There were no students designated as English Language Learners (ELL) in the dataset. To mediate the potentially confounding influence of the extended constructed response items exclusively within scenario question types, a further comparison was made by examining performance on multiple-choice items only within scenario and stand-alone question types.

Summary of Results

An examination of performance of eighth-graders on the 2005 science WASL uncovered some interesting results for gender, ethnicity, and question type. The summary of the findings are presented for overall performance by gender and ethnicity and for question type by gender and ethnicity.

In terms of overall performance on the 2005 WASL science exam, girls' mean score was slightly higher than boys'; however, this difference was not statistically significant. The percentage of boys and girls that met proficiency was similar (38% and 37%, respectively) as was the distribution of scores within each proficiency level. Within each ethnicity, girls and boys performed similarly.

On the overall test, significant differences were found between ethnic subgroups. The differences were similar to the achievement gap found nationally in the United States. Approximately 40% of the students in the Multiracial, White, and Asian subgroups met proficiency. In contrast, only 20% or less of the students in the American Indian/Alaskan Native, Black African American, Hispanic, and Hawaiian/Pacific Islander subgroups met proficiency. A majority of the non-proficient students in the lower-scoring subgroups were in the lowest proficiency level (Level 1 - Below Basic).

Girls' mean score was slightly higher than boys' mean scores on both stand-alone and scenario question types; however, when multiple-choice items only were considered, girls' mean scores dropped considerably below boys' mean score on both question types. Mean scores within gender varied little between question type.

In general, all subgroups except Asians and Whites scored higher on stand-alone versus scenario question types. The largest increases in mean score on stand-alone question types compared to the mean score on scenario question types was realized by the lowest scoring subgroups. The same patterns of performance emerged with and without the constructed response items included in the analysis.

Previous studies suggest that performance-like tests that provide an authentic context for questions can facilitate students' comprehension and that this may result in higher achievement scores within underrepresented groups (Stecher et al., 2000). However, this study found that within the science domain, presenting test items in an authentic, performance-like context did not help to reduce the science achievement gap. Possible explanations for this finding will be discussed later in this chapter.

Gender & Ethnic Differences on the Overall WASL Test

This study set out to examine ethnic and gender group differences on the overall 2005 eighth-grade science WASL. Eighth-grade boys and girls in Washington State performed similarly on the WASL science exam and the distribution of scores within each proficiency level were similar. Girls' mean scores were slightly higher than boys' mean scores but the difference was not significant. Prior research (e.g., Van-Largen, Bosker, & Dekkers, 2006) indicated superior performance of boys compared to girls in science achievement, but the results of this study did not support this finding. There was no gender gap in achievement scores on the 2005 eighth-grade science WASL. Several factors may have contributed to the similarities in achievement scores.

One factor that may help to explain this outcome is the inclusion of a variety of item response formats rather than the exclusively multiple-choice response format. The WASL assessment design uses multiple measures and multiple response formats that may help to create a gender-fair assessment tool. Previous studies have found that boys' advantage is smaller on essay items compared to multiple-choice items (Mazzae, Schmitt & Bleinstein, 1993; Zenisky, Hambleton & Robin, 2004). Multiple-choice response formats favor students who guess (Rowley, 1974) and male students tend to guess more than female students do (Hanna, 1986). Washington State's Science Standards follow the National Science Education Standards (NRC, 1996) and emphasize science as inquiry and the development of abilities to do science. Hence, the science WASL uses alternative formats, such as extended constructed response, for students to demonstrate their ability to conduct safe experiments, construct and interpret data tables as well as other performance-like attributes that characterize the science domain. The literature suggests gender differences can be reduced on items that involve multistep reasoning, designing experiments, and drawing conclusions (Erickson & Erickson, 1984; Linn, Benedictis, Delucchi, Harris, & Stage, 1987) possibly because these items place a greater emphasis on critical thinking and less reliance on practical science experiences outside the classroom (Jovanovic et al., 1994).

Historically, boys tend to score significantly higher than girls on test items within the physical and earth science content areas compared to test items within the biology and life science content areas (Becker, 1989; Erickson & Erickson, 1984). In addition, boys tend to score much higher than girls on practical items compared to theoretical items (Sencar & Eryilmaz, 2004). Researchers suggest that differences in background or out-of-school experiences may contribute to these differences in science achievement scores (Jovanic et al., 1994; Johnson, 1987). However, the 2005 eighth-grade science WASL included a balance of items within the physical, earth, space, and living systems content areas as well as a balance of practical and theoretical items. This balance may have mitigated the effect of gender found in previous studies.

On overall test performance, there were significant differences in achievement between ethnic subgroups. In the top three performing ethnic subgroups (Multiracial, White, Asian), approximately 40% of the students were proficient. This figure drops substantially in the remaining subgroups where approximately 20% or less of the students were proficient. Furthermore, most of the non-proficient students in the remaining subgroups also score in the lowest proficiency level. Differences in achievement scores within ethnic subgroups continue to be reflected in test performance not only in Washington State but also on a national level (Department of Education, 2005). For example, 2005 NAEP scores show the gap between White and Black or White and Hispanic students as the largest among ethnic groups.

Differential Item Functioning (DIF) is a statistical method used to examine response patterns across ethnicity, gender and language groups or “cultural validity”

(Solano-Flores, 2002). The results from this analysis conducted by OSPI, indicated that the 2005 science WASL was a fair assessment for measuring the knowledge of underrepresented groups. However, a clearer picture of cultural validity would provide insight into why a student may have answered an item incorrectly, which can be different for each student (National Research Council, 2001).

One promising model for addressing this problem is Bayesian Inference Networks (BINs). This model permits differentiation between sources of item difficulty, such as language (Erikan, 1998) or solution strategies (Lane, Wang, and Magone, 1996) which may help to reduce achievement differences between ethnic subgroups. BINs extend the principles of test theory and model the multidimensional nature of student problem solving. The model captures the interrelationships between large numbers of variables within small subsets and can take into account personal variation in fine-grained skills used in problem solving (Pardos, Heffernan, Anderson & Heffernan, 2006). This means that the cultural validity of tests may improve because sources of item difficulty can be isolated. When sources of item difficulty are isolated, issues of language can be handled differently than, for example, issues of insufficient domain knowledge.

Currently, the use of BINs requires a strong knowledge in statistical modeling; however, in the future, this modeling technique may be incorporated into user-friendly software. BINs have the potential of being a useful tool that can inform classroom instruction (NRC, 2001). For example, information on strategy use may be captured as students complete questions administered by the software program. Then, students' responses are fit to a model that best represents where they are along the novice-expert continuum. Finally, both student and teacher would receive information that pinpoints where a student is along their scaffold of learning on many dimensions, including areas that are not relevant to the measured construct.

Gender & Ethnic Differences on Scenario versus Stand-Alone Question Types

The study also set out to compare the performance of gender and ethnic subgroups on scenario question types versus their performance on stand-alone question types. Within gender, mean scores varied little between question type. Girls' mean score was higher than boys' mean score on both stand-alone and scenario question types; however, this difference was not statistically significant.

In order to determine if item response format contributed to gender fairness, constructed response items were removed from the analysis. The changes in students' performance were striking. First, mean scores on both question types decreased for each gender. Second, boys' mean scores were higher than girls' mean scores for both question types. This finding is interpreted to mean that the score advantage girls' had on the WASL was not present when only multiple-choice items were considered. This interpretation is supported by previous research that suggests boys tend to score higher than girls on multiple-choice items (Ben-Shakhar & Sinai, 1991) and girls tend to score higher than boys on performance-like items (Janovic et al., 1994).

Linn (1992) and Klein et al. (1997) found that females perform better than males on constructed response and essay formats designed to measure integrated understanding compared to multiple-choice response formats that measure general science phenomena.

Scenario question types measure integrated understanding using items with an extended constructed response (ECR) format. ECR items require more than two steps and longer (< 3 sentences) responses than short constructed response items found in stand-alone question types. ECR items may ask students to develop a viable solution, demonstrate an understanding or process, communicate ideas or results, or show reasoning using complex responses. They may also ask for a graph, figure, diagram, and/or table with labels, words, sentences, or equations to support students' reasoning. When ECR items were removed from the analysis, girls' achievement scores dropped substantially below boys' achievement scores, suggesting the inclusion of this item format results in a more gender-fair assessment.

With respect to ethnicity, performances between ethnic subgroups were significantly different on both scenario as well as stand-alone question types. The rank order of the four highest scoring ethnic subgroups was the same regardless of question type. From highest to lowest, the four highest scoring subgroups were White, Multiracial, Asian, and American Indian/Alaska Native. The three lowest scoring subgroups were Hawaiian/Pacific Islander, Black, and Hispanic.

To determine if one question type was better for students in underrepresented groups to provide evidence of learning, achievement scores on both question types were compared within each ethnicity. Hispanic, Black, and American Indian/Alaskan Native students performed much better on stand-alone question types compared to scenario question types. The largest point gains on stand-alone question types were made by the lowest performing subgroups. Thus, the point gap between the highest and lowest performing groups decreased on stand-alone question types.

Performance within each ethnic subgroup between question type was further examined by calculating the difference (in standard deviation units) between scenario and stand-alone T-score points. A method called binomial effect size display (BESD) (Rosenthal & Rubin, 1982) was used to contextualize and interpret the magnitude of achievement differences between scenario and stand-alone question types for each ethnic group. For some ethnic groups, using stand-alone question types would mean substantial number of students with improved scores. Using BESD, the effect size difference in the American Indian/Alaskan Native and Hispanic subgroups can be interpreted as approximately 8% more American Indian/Alaska Native (175) and 9% more Hispanic (784) students scored higher on stand-alone compared to scenario question types. Approximately 5% more Black (216) students and approximately 2% more Multiracial (5) students scored higher on stand-alone compared to scenario question types. On the other hand, Asian, Hawaiian Pacific Islander, and White students favored scenario question types. Using BESD, the magnitude of achievement difference between stand-alone and scenario question types can be interpreted as approximately 5% more Asian (302) and Hawaiian Pacific Islander (3) students and 2% more White (1,122) students achieved higher scores on scenario compared to stand-alone question types. Thus, the simultaneous increase in mean score on stand-alone question types for lower performing subgroups and decrease in mean score on stand-alone question types for higher performing subgroups had the effect of reducing the achievement gap.

These results suggest that scenario question types discriminate more between low and high performing students. Some items exclusively within scenario question types require the student to construct lengthy, multistep explanations, patterned after NAEP performance-like assessments. Research on NAEP items concluded that students did well on simple, straightforward items but had more difficulty with extended constructed response items (Jones, Mullis, Raizen, Weiss, & Weston, 1992). Within the science WASL, scenario question types require students read, understand and extract information in the preceding paragraph of text in order to answer the questions correctly. Scores on scenario question types possibly reflect students' reading ability as well as science domain knowledge. When examining the performance assessment scores on science inquiry items of eighth-grade students, Stecher et al. (2000) similarly concluded that science inquiry items rely heavily on students' reading ability.

Extended constructed response items within scenario question types require students to work with more than one piece of information at a time to answer the multistep item. Performance on NAEP multistep items is typically lower than performance on single step items due to omission of one part of a two-part item (Jones et al., 1992). OSPI has found a similar error pattern on past science WASL exams, which led to the addition of the bolded sentence stem in the center of the student response page (R. Beven, personal communication, December 18, 2006). The intention of the bolded sentence stem is to serve as a reminder to complete the second part of the item. Issues such as the mechanics of test-taking serve as a reminder that every assessment includes a measurement of the degree a learner can participate in the practice of test-taking (NRC, 2001).

Explanations for Achievement Differences From Prior Research

Three prevailing themes emerge from the literature as plausible explanations for achievement differences between groups: individual differences (cognitive theory), opportunity to learn (OTL), and attributes of the test items (Table 16). The first theme, individual cognitive differences, is informed by empirical and theoretical studies in cognitive psychology that support individual differences in performance and learning. The expert/novice model (Chi et al., 1988; Alexander, 2003) is used to frame this discussion. There are many factors in the second theme, opportunity to learn (OTL), that are linked to achievement differences (Darling-Hammond, 2000). This discussion examines two OTL factors, quality teaching and the enacted curriculum (Yoon & Resnick, 1998). The third theme, attributes of test items, has been demonstrated to affect student performance on achievement tests. Item attributes, such as instructional sensitivity (D'Agostino et al., 2007) and passage length (Davies, 1988) are discussed.

Table 16. Themes and Factors from Prior Research Demonstrated to Influence Student Performance on Achievement Tests.

I. Individual Differences (Cognitive Theory)

- Expert/Novice theory (Chi et al, 1988; Alexander, 2003)
- Cognitive components (D'Andrade, 1992; Vosniadou, 1994)
- Cognitive correlates (Anderson, 1982; de Ribaupierre & Rieben, 1995)

II. Opportunity to Learn

- School context characteristics (Roza et al., 2007)
- Quality teaching and learning (Darling-Hammond, 2000)
- Classroom context variables (Colker et al., 2003)
    - Intended curriculum and instruction (Schmidt, et al. 1997)
- Classroom process variables (Colker et al., 2003)
    - Enacted curriculum and instruction (Yoon & Resnick, 1998)

Table 16. Themes and Factors from Prior Research Demonstrated to Influence Student Performance on Achievement Tests (continued).

III. Attributes of Test Items

- Instructional sensitivity (D'Agostino et al., 2007; Ruiz-Primo et al., 2002)
- Passage length (Davies, 1988)
- Academic vocabulary (Shaftel et al., 2006)
- Cognitive complexity (Klein et al., 1997)
- Science strand type (Bruschi & Anderson, 1994)
- Degree of knowledge transfer (Alexander & Judy, 2003)

Theme I – Individual Differences

The study of individual differences is divided into two approaches: 1) cognitive correlates and 2) cognitive components (Glaser & Pelligrino, 1979). The cognitive components research attempts to link differences in student test performance to cognitive components such as memory span, spatial visualization, and inductive reasoning. Individual differences have also been linked to cognitive components, such as mental models (Glaser & Baxter, 1999; Vosniadou, 1994) and schemas (D'Andrade, 1992). The cognitive correlates approach seeks to explain differences in test performance through the mechanics of information processing such as speed, executive routines, and efficiency. Individual differences have been linked to cognitive correlates such as working memory processing speed (Woltz, 2003) and modes of processing (de Ribaupierre & Rieben, 1995; Reuchlin, 1978). Further, expert/novice theory suggests domain and problem-solving schemas deepen qualitatively and increases quantitatively as an individual moves from a level of acclimation to one of proficiency (Alexander, 2003). This facilitates quicker information processing and retrieval of information (Chi et al., 1988).

The largest gaps between high and low performing subgroups were found on scenario question types which may require more developed domain and problem-solving schemas compared to stand-alone question types. The performance-like attributes of scenario question types lend themselves to assess constructs within the Inquiry and Design Science Strands. Mastery of this standard requires the student to transfer scientific processes, procedures, and principles to a new investigation with a different manipulated and/or responding variable. Appendix B includes an example of a scenario question type called “In the Doghouse.” This WASL item requires students to demonstrate the following evidence of learning: 1) the formulation of a hypothesis, 2) reasons that support the prediction, 3) identification of a controlled variable, 4) identification of a manipulated (independent) variable, 5) identification of a responding (dependent) variable, 6) creation of a logical step-by-step plan for a similar investigation, 7) description of materials and tools needed, and 8) organization of data tables from multiple trials. The logical systematic plan developed by the student for the new investigation must have clear, repeatable steps. An assessment item such as this is considered a test of far transfer as opposed to near or zero transfer because the context is novel and new to the student (NRC, 2001).

Successful application of knowledge to new contexts is facilitated by practice and is one of the traits of a content area expert (Chi et al., 1988). However, in the novice stages of learning, performance is heavily dependent on limitations of working memory (Anderson, 1982). When the skill is exercised frequently, the skill becomes fluent, freeing up working memory to better focus on the meaning and execution of the problem (Anderson, 1982).

Strategic knowledge plays a strong role in the acquisition of domain knowledge and can help guide students to surmise information from example problems and information that is readily accessible (Alexander, 2003). Some novice students have more developed strategic knowledge and better metacognitive skills than other students (Chi et al., 1988). Indeed, models within the study of human development acknowledge that major differences exist both within and across individuals in the development of cognitive processes used to solve mathematics and science tasks (de Ribaupierre & Rieben, 1995). These processes, called propositional processes, develop as students get older but students differ in their flexibility to apply different modes of processing (de Ribaupierre & Rieben, 1995), which may compromise achievement on mathematics and science tasks.

Earlier research has also indicated large differences between White and Black/African American or Hispanic students on items within NAEP's Nature of Science Strand (Bruschi & Anderson, 1994). Like Washington State's Inquiry of Science Strand, the NAEP Nature of Science Strand assesses understanding of the scientific processes including the formulation of a prediction, design of an experiment, inference and interpretation of the data.

Cognitive theory states that knowledge frequently develops in a highly contextualized form indicating that the best retrieval cues for the knowledge are the same as the learning cues for the knowledge (Thomas & Tulving, 1970; Tulving & Thomson, 1973). For example, Bilsky, Blachman, Chi, Mui, and Winter (1986) found that, when target items are embedded within a series of story passages, comprehension is facilitated only when the story passages guide the selection of a relevant problem-solving schema.

Recall and recognition are the two methods used to retrieve information from long-term memory. Recall refers to information retrieval that is non-cued such as an assessment item that asks, “How can Dahlia change the entrance to the dog house so that more heat energy is retained and why does the design work?” Recognition, on the other hand, refers to information retrieval that is assisted by cues such as diagrams and bolded sentence stems embedded in the assessment item. Recognition occurs within two systems of memory representation: the verbal and imaginal (or mental image) systems. The more cues that are used in encoding, the more likely one or another of them will be available to facilitate retrieval (Driscoll, 2000). This suggests that classroom instruction that focused on student acquisition and retrieval of learning would present the standard in a similar manner as the assessment tool. This approach is also supported by schema theory (D’Andrade, 1992). Schema-based research in educational psychology and education measurement fields emphasizes the importance of providing a coherent structure (schema) for the organization of knowledge that can be used as a framework for curriculum and assessment (Marshall, 1995). Teachers could target essential schemas, based on the state standards, for classroom instruction. Thus, alignment would be achieved between classroom instruction, state standards, and the assessment tool. Studies have indicated that achievement scores in all domains except social studies are related to how well the standards, classroom instruction, and the state standardized exam are aligned (Borko & Stecher, 2001; D’Agostino, Welsh, & Corson, 2007).

Theme II - Opportunities to Learn

The development of knowledge and skills in the science domain within a classroom context includes specific forms of practice such as inquiry, data collection, hypothesis testing, and the transfer of theory to new applications. Inferences about learning may not be accurate if students have had limited or non-existent opportunities in the classroom to participate in forms of practice required for success in the science domain. Achievement scores may in part reflect the differences in opportunities to learn (OTL) (Kher, Schmidt, Houang, & Zou, 2007). This term has evolved to encompass many more variables than the original definition, which focused only on classroom context variables such as the intended curriculum and actual instruction (Colker, Toyama, Trevisan, and Haertel, 2003). Colker et al. (2003) provided a useful framework to categorize variables currently suggested by the OTL literature. They grouped variables into system context variables, (such as levels of funding), school context variables, (such as course offerings), teacher background characteristics, classroom context variables, (such as intended curriculum and instruction), and finally, classroom process variables, (such as the enacted curriculum and instruction). For the purposes of this research, teacher background characteristics and the enacted curriculum and instruction will be discussed as they may help explain the variance in performance between ethnic subgroups and lower achievement scores on scenario versus stand-alone question types.

Quality Teaching and Learning. Differences in the quality of teaching and learning have been linked to large disparities in student achievement between ethnicities (Darling-Hammond, 2000; Zurawski, 2004). Quality teaching and learning is one of several school attributes that support students in attaining standards (Shannon & Bylsma, 2007); however, frequently teachers with less experience, less education, and less content expertise are assigned to the schools in most need (Peske & Haycock, 2006). Schools with high minority and/or low-income enrollment are not only vulnerable to staffing inequities but also to funding inequities (Roza, Guin, & Davis, 2007).

The most effective schools use curricular and organizational strategies such as collaborative learning structures for both students and teachers that support the alignment of curriculum, instruction and assessment (Shannon & Bylsma, 2007). Schools with these features can provide teachers with professional development opportunities, such as embedded professional development designed to improve classroom instruction and support students in attaining the standards.

A central tenet of the Washington State science standards is that a proficient student can "do science" demonstrated by the ability to support claims, record and analyze data, and draw sound conclusions. These activities are consistently correlated to a deep understanding of the nature of science (Sinclair, 1994; Von Secker & Lissitz, 1999) and often require direct instruction (Garner, 1987). High-need schools may lack support, supplies, and equipment to provide the practice and scientific discourse that leads to this deeper understanding of the nature of science and science inquiry (Rodriguez, 1997).

The creation of learner-centered educational experiences is particularly challenging because teachers in high-need schools may lack the professional development needed to shift from teacher-centered to learner-centered instruction

(Lambert & McCombs, 1998). The opportunity for students to interact supportively with one another has also been linked to gains in students' ability to communicate and reason mathematically (Ginsburg-Block & Fantuzzo, 1998), but teachers may lack the training needed to successfully mediate social discourse as learners actively construct knowledge within the science domain (Driver, Asoko, Leach, Mortimer & Scott, 1994). Therefore, students' development of critical thinking skills may be compromised in high-need schools where learner-centered instruction, which has been linked to large science achievement gains in all ethnic subgroups (Von Secker & Lissitz, 1999), is not practiced.

Enacted Curriculum and Instruction. With the advent of diploma-sanctioned testing in this country, OTL concepts were applied to the enacted curriculum and instruction within the classroom (Colker et al., 2003). In other words, if decisions about high school graduation to be based on state standardized test scores, then the student must have been exposed to the knowledge the test is designed to measure. This focus grew out of concern over the fairness of basing decisions of high school graduation on state standardized test scores, if the student was not exposed to the knowledge the test was designed to measure. Instructional validity received much attention following the court case of Debra P. v. Turlington (Phillips, 1993) which established that diploma sanctioned testing must have instructional validity.

Evaluation of instructional validity is complex. One approach advocates that if the state standards are used to develop the state standardized exam, then that is sufficient evidence of instructional validity. Indeed, The Committee on the Foundations of Assessment recommends that assessment development begin with the model of cognition, as represented by the state standards (NRC, 2001). This approach drives alignment between the model of cognition (standards) and observations designed to demonstrate evidence of learning (assessment tool).

An alternative method of the evaluation of instructional validity is to examine instruction at the classroom level. If classroom instruction is aligned to state content standards, then instructional validity can be inferred. Airasian & Madaus (1983), succinctly defined construct validity in the following statement:

> "If one were to question whether the processes underlying performance on the test are, in fact, related to the specific instruction received, as opposed to other non-school or program factors (e.g., general pupil ability, social class), one would be questioning the construct validity of the test and thus its usefulness for making inferences about school or program effectiveness." (p. 106)

Critics of this approach point out the infeasibility of examining instruction in every classroom but some research suggests that construct validity cannot be determined simply by an inspection of test items (D'Agostino et al., 2007).

Several studies have examined the enacted curriculum and instruction at the classroom level to determine how varying types of instructional strategies and approaches can produce different patterns of achievement. Reform-based teaching practices, such as inquiry-based practices, are consistently related to higher student achievement (Cohen & Hill, 2000; Ginsburg-Block & Fantuzzo, 1998; Von Secker, 2002).

The strength of alignment between instruction and assessment was demonstrated in a study where teachers used a consistent deliberate instructional method designed to assist students in transferring their knowledge to new applications (Bottge, 1999). Specifically, Montague's (1997) cognitive strategy training model was used to show students how to recognize common patterns in problems and purposefully match the pattern to one previously encountered. Montague's model emphasizes the process of students verbally paraphrasing the problem and stating a prediction, followed by an estimation, computation, and check for each problem. Not only does this process aid in the transfer of skills to new applications but also frees up working memory for planning and problem execution (Cooper & Sweller, 1987). After the contextualized and cognitive strategy instruction, both low and medium achieving students had significantly higher scores on a test that assessed the transfer of skills to new applications compared to the group that did not receive the specialized instruction.

Theme III - Attributes of Test Items

From an educational policy perspective, current federally mandated policy requires schools to report Annual Yearly Progress (AYP), implying that state tests are a measure of school effectiveness. Inferences about school effectiveness, including curriculum and instruction, are made based on state exam achievement scores (Hattie et al., 1999). Some researchers urge that if decisions regarding the effectiveness of schools are made based on achievement scores, then the scores must be closely linked to and influenced by instruction (Airasian & Madaus, 1983; Cronback & Meehl, 1955; D'Agnostino et al., 2007; Haydel, 2003; NRC, 2001; Webb, 1997).

Instructional sensitivity, a narrower concept of instructional validity, focuses on the assessment items themselves. This term refers to the sensitivity of test items to register differences in classroom instruction (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002; Yoon & Resnick, 1998). Instructional sensitivity of test items is deemed important in the standards-based education reform movement because teachers' commitment to improve student learning by teaching the standards may diminish if the testing instruments fail to register their efforts (D'Agostino et al., 2007). Indeed, performance on test items should improve with teaching (Baker, 1994).

Some content domains and grade levels tend to be more instructionally sensitive than others (Airasian & Madaus, 1983). Domains that test school specific skills, such as mathematics and science, have a greater degree of instructional sensitivity than the reading domain. Subjects such as reading typically include more variation due to home environment differences (Madaus, Airasian, & Kellaghan, 1980). Further, some test items may be more instructionally sensitive than others, but aggregate test scores can mask the sensitivity of each test item (Airasian & Madaus, 1983). It is plausible that items within the scenario question type are more instructionally sensitive than stand-alone items.

D'Agostino et al. (2007) explored the instructional sensitivity of the fifth grade mathematics Arizona Instrument to Measure Standards (AIMS) test. The items on this test were designed to measure students' understanding of published state mathematics standards. The study investigated the alignment between detailed descriptions of evidence of learning used to create test items and teachers' classroom instruction. For example, one measure of alignment was the match between academic vocabulary and learning objectives published in the standards and academic vocabulary and learning objectives used by the teachers in the classroom. The amount of instructional time, or emphasis, dedicated to the standard was also measured. Out of the two measured variables, alignment and emphasis, alignment explained most of the achievement differences between classrooms. Emphasis, or the amount of instructional time, did not matter if the instruction was not aligned with the learning objectives in the standards. The highest mean achievement scores were realized in classrooms where teachers not only covered the content but also presented the information the same way it was tested. The authors concluded that items on the AIMS test are sensitive to differences in instruction.

In another study, Cohen and Hill (2000) examined the instructional sensitivity of the California Learning Assessment System (CLAS) mathematics test. Designers of the newly revised state test were optimistic that the assessment items focused on reform-based standards and would be sensitive to teacher instruction in these standards. They found a positive relationship between teacher-reported frequency of the use of reform-based teaching practices and achievement scores on the CLAS mathematics test at the school level, after controlling for demographic characteristics. In contrast, Mayer (1998) found trivial or no relationship between teacher-reported frequency of the use of reform-based teaching practices and achievement scores on dated, pre-reform traditional multiple-choice tests.

Peak (1996) commented that in today's standards-based reform movement, teachers have an awareness of the standards, but their approach to classroom instruction varies and is not always aligned to the standards. If students are not receiving classroom instruction that is aligned with the standards, there may be greater ambiguity and complexity with scenarios that assess science inquiry and application skills. Students have remarked that performance-like assessments are “fun” and support learning more than decontextualized multiple-choice items, but they are also uncertain how to answer and how their answers will be graded (Haydel, 2003). Students’ performance improves when classroom activities provide practice in the application of problem-solving schemas similar to schemas needed to demonstrate proficiency on assessments (Anderson, 1990).

Well-intended performance-like scenarios may in fact be a barrier to comprehension due to the characteristics of the text passage. Scenario question types require students to read a lengthy text passage that describes the authentic context for the target items that follow. Features of question stems, such as the number of words, have been found to increase the difficulty of reading comprehension for readers of all ability levels (Davies, 1988).

The literature on the theory of settings may also help to explain why the findings from this study diverged from the expected outcomes. The theory of settings suggests that when the environment varies, a student’s application of strategies varies as well (Garner, 1990). It is possible that students respond differently in a high-stakes testing environment compared to a low-stakes testing environment where most of the previous research exploring the role of contextual items on test performance has been conducted. Distinctly different factors may be at play in these two environments and inferences made about student performance in one environment may not be valid or readily transferable to another environment (Garner, 1990).

Policy Implications

The reasons that underrepresented groups earned fewer score points on scenario versus stand-alone question types remains unclear. Student performance on large-scale tests is the result of a complex interaction of factors that we do not fully understand and the scope of this study did not permit an in-depth analysis of all of the factor involved. Hence, it is necessary to be cautious in drawing implications about the effect of question type or item response format on student performance on the science WASL. Designers of assessments and policy makers may be tempted to include more stand-alone multiple-choice items on assessments; however, this decision is risky for three reasons.

First, national and international policies on science education place great value not only on knowledge within the science domain but the ability to apply science knowledge to solve human problems (National Academy of Science, 1996; TIMMS, 2004). One challenge of assessment design is to construct items that effectively measure the problem solving and critical thinking skills published in the National Science Education Standards. Scenario question types lend themselves to the demonstration of evidence of learning these skills and may not be decomposable into smaller pieces. Shorter, stand-alone question types have fewer words to decode and may not be as instructionally sensitive as scenario question types, but this design may be at the expense of ignoring science values demonstrated to be important both nationally and internationally.

Second, Washington State's unique approach to the design of the science WASL creates an assessment instrument aligned to the state standards. Task construction should begin with the chosen model of cognition (standards) (NRC, 2001). This will help to create assessment items that are tightly linked to the standards and tap into the intended construct. This alignment is important because efforts to close both the testing and learning gap are compromised when the assessment instrument is misaligned with the standards (Li, Klahr, & Siler, 2006).

Third, it is critical that the design of assessments yield fair and valid inferences about students' knowledge. This research suggests that scenario question types that included both constructed response and multiple-choice items were most effective for girls to show evidence of learning. Adoption of an assessment with only multiple-choice items quite possibly could increase the gender gap and reduce the fairness of the science WASL.

In sum, the debate is not whether to abandon scenario question types but rather in the near term, how to improve scenario question types. In the long run, it is about investing in a more balanced assessment system. This research reiterates to policy makers that important decisions about students and their learning need to be based on more than a single test score (NRC, 2001). Important decisions should be based on multiple-measures in varying formats, such as end-of-course grades and student portfolios.

The educational context in which students learn is important when interpreting student achievement scores (Garden et al., 2006). The performance gap between ethnic subgroups on scenario question types may be an indication of compromised opportunities to learn within some districts in Washington State. To address this problem, policy makers can promote a more equitable balance of resources to high-need schools. This would provide the organizational and curricular support needed for students and ongoing professional development in science education for teachers.

Teachers need to increase their familiarity with Washington State's science standards and reflect on how they are brought to life within the classroom. Each standard has a section entitled "Classroom Evidence" that provides detailed information about the skill set needed to achieve proficiency. This detailed skill set provides the opportunity to align lessons to standards and bring continuity between classroom formative assessments and the WASL. Released WASL items can provide time for students to practice and refine problem-solving strategies. Teachers can engage in classroom activities that encourage self-assessment and foster metacognitive skills in students, such as examining grading rubrics and exemplar work.

Conclusions and Suggestions for Future Research

This study started with a focus on student performance as related to question type, namely, scenario and stand-alone question types. Based on prior research, it was predicted that students might realize higher achievement scores on authentic, contextualized items within scenario question types compared to less contextualized stand-alone question types. The results of this research, however, demonstrated that students earned more points on stand-alone compared to scenario question types. In an effort to explain these results, it was surmised that the unique characteristics of certain items exclusively within the scenario question type might be contributing to the significant achievement differences between ethnic subgroups. Given that total test scores can mask item differences, future research needs to focus on explaining variance at the item level.

Tests within the science domain have a tendency to be sensitive to variations in classroom instruction. Future research is needed to evaluate the instructional sensitivity of items that may affect students' performance on the science WASL and inadvertently disadvantage some students. Ideally, the standards-based reform movement encourages teachers to focus on teaching the specified content. Herman (2004) commented, "The idea is not really to teach to the test, but to motivate everyone in the system to focus on the standards and enable children to reach them" (p. 143). Future research needs to explore more precisely the degree to which high and low achievement scores on items are related to classroom instruction. A better understanding is needed about how standards affect classroom instruction and the degree of alignment between what is being taught and what is being assessed.

A closer examination of students' work may uncover particular error patterns that can be addressed through the improvement of assessment design. For example, error patterns (such as answering only one part of a two-part question) and omission rate can sometimes be remedied by changing the font to a larger, bolder size or providing more white space in the item design (Sugrue, 2001). In addition, error patterns may reveal that students' misinterpret key details of the scenario question types. It is possible that variance in achievement scores may be exacerbated by less developed academic English language proficiency. Academic English schemas may not be well developed and therefore not cued.

This study used MANOVA for statistical analysis of data and was limited to the extracted data set provided by OSPI. Future research could employ a statistical modeling technique, such as Hierarchical Linear Modeling (HLM), and additional nested data from the state assessment database may provide new insights. This form of analysis could determine if the same types of differences found in this study exist within question type by schools within districts or within question type by classrooms within schools. This information could help guide the creation of professional development programs targeted to provide organizational and curricular support so that all students have equal opportunities to learn and master Washington State standards within the science domain.

REFERENCES CITED

Ahmed, A., & Pollitt, A. (2007). Improving the quality of contextualized questions: an experimental investigation of focus. Assessment in Education, 14(2), 201-232.

Airasian, P. W., & Madaus, G. F. (1983). Linking testing and instruction: Policy issues. Journal of Educational Measurement, 20(103-117).

Alexander, P. A. (1997). Mapping the multidimensional nature of domain learning: the interplay of cognitive, motivational, and strategic forces. In M. L. Maehr & P. R. Pintrich (Eds.), Advances in motivation and achievement (Vol. 10, pp. 213-250). Greenwich, CT.: JAI Press.

Alexander, P. A. (2003). The development of expertise: The journey from acclimation to proficiency. Educational Researcher, 32(8), 10-14.

Alexander, P. A., & Judy, J. E. (1988). The interaction of domain-specific and strategic knowledge in academic performance. Review of Educational Research, 58(4), 375-404.

Allard, F., & Starkes, J. L. (1991). Motor-skill experts in sports, dance and other domains. In K. A. Ericsson & J. Smith (Eds.), Toward a General Theory of Expertise: Prospects and Limits (pp. 126-152). Cambridge: Cambridge University Press.

Allen, M. J., & Yen, W. M. (1979). Introduction to Measurement Theory. Monterey: Brooks/Cole Publishing Company.

Anastasi, A., & Urbina, S. (1997). Psychological Testing. Delhi: Pearson Education.

Anderson, J. R. (1982). Acquisition of a cognitive skill. Psychological Review, 89(4), 369-406.

Anderson, J. R. (1990). Cognitive Psychology and Its Implications (3rd ed.). New York: W.H. Freeman.

Ausubel, D. P. (1960). The use of advance organizers in the learning and retention of meaningful material. Journal of Educational Psychology, 51, 267-272.

Ayers, L. P. (1918). History and present status of educational measurements. In G. M. Whipple (Ed.), The measurement of educational products (Vol. Seventeenth yearbook of the National Society for the Study of Education, Part II, pp. 9-15). Bloomington, Ill.: Public School Publishing Company.

Baghetto. (2007). Factors associated with middle and secondary perceived science competence. Journal of Research in Science Teaching, 44(6), 800-814.

Baker, E. L. (1994). Issues in policy, assessment and equity. In B. Farr & E. Trumbull (Eds.), Assessment Alternatives for Diverse Classrooms (Vol. 2, pp. 1-18). Norwood, MA.: Christopher-Gordon Publishers, Inc.

Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1992). Evaluation of a procedure-based scoring for hands-on science assessment. Journal of Educational Measurement, 29(1), 1-17.

Bechtel, G. A., Davidhizar, R., & Bunting, S. (2000). Triangulation research among culturally diverse populations. Journal of Allied Health, 29(2), 61-63.

Becker, B. J. (1989). Gender and science achievement: A reanalysis of studies from two metaanalyses. Journal of Research in Science Teaching, 26, 141-169.

Bennett, R. E., & Ward, W. C. (1993). Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment. Hillsdale, NJ: Lawrence Erlbaum.

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. Journal of Educational Measurement, 27(2), 165-174.

Bideaud, J. (1988). Analogical processing and propositional processing in the resolution of a Piagetian task. Archives of Psychology, 56, 295-300.

Biggs, J. B., & Collins, K. F. (1982). Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome). New York: Academic Press.

Biggs, J. B., & Collins, K. F. (1989). Toward a model of school-based curriculum development and assessment using the SOLO taxonomy. Australian Journal of Education, 33, 151-163.

Bilsky, L. H., Blachman, S., Chi, C., Mui, A. C., & Winter, P. (1986). Comprehension strategies in math problem and story contexts. Cognition and Instruction, 3(2), 109-126.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Educational Assessment: Principles, Policy and Practice, 5(1), 7-74.

Bond, L., & Glaser, R. (1979). ATI, but mostly A and T and not much of I. Applied Psychological Measurement, 3, 137-140.

Borko, H., & Stecher, B. M. (2001). Looking at reform through different methodological lenses: Survey and case studies of the Washington State education reform. Paper presented at the American Educational Research Association, Seattle.

Bottge, B. A. (1999). Effects of contextualized math instruction on problem solving of average and below-average achieving students. The Journal of Special Education, 33(2), 81-92.

Bradsford, J. D., Brown, A. L., & Cocking, R. R. (1999). How People Learn: Brain, Mind, Experience, and School Committee on Developments in the Science of Learning. Washington, D. C.: National Academies Press.

Bransford, J. D., & Stein, B. S. (1984). The ideal problem solver: A guide for improving thinking, learning, and creativity. New York: W.H. Freeman.

Broadfoot, P. B., P. (2004). Redefining assessment? The first ten years of Assessment in Education. Assessment in Education, 11(1), 7-27.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. Educational Researcher, 18(1), 32-42.

Bruschi, B. A., & Anderson, B. T. (1994). Gender and ethnic differences in science achievement of nine-, thirteen-, and seventeen-year-old students. Paper presented at the Annual Meeting of the Eastern Educational Research Association, Sarasota, FL.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. Cognitive Psychology, 1, 33-81.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic versus syntactic approaches to training deductive reasoning. Cognitive Psychology, 17, 391-416.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5, 121-152.

Chi, M. T., Glaser, R., & Farr, M. J. (1988). The Nature of Expertise. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ciofalo, J., & Wylie, C. (2006). Using diagnostic classroom assessment: One question at a time (No. 12285). Princeton: Educational Testing Service.

Cohen, D. K., & Hill, H. C. (2000). Instructional policy and classroom performance: The mathematics reform in California. Teachers College Record, 102(2), 294-343.

Cole, M. (1996). Cultural Psychology: A once and future discipline. Cambridge: Harvard University Press.

Cole, N. S., & Moss, P. A. (1989). Bias in Test Use. In R. L. Linn (Ed.), Educational Measurement (pp. 201-219). New York: American Council on Education and MacMillan.

Colker, A. M., Toyama, Y., Trevisan, M., & Haertel, G. (2003). Literature review of instructional sensitivity and opportunity to learn studies. Paper presented at the Annual meeting of the National Council on Measurement in Education, Chicago.

Cooley, W. W., & Leinhardt, G. (1980). The instructional dimensions study. Education Evaluation and Policy Analysis, 2, 7-25.

Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. Journal of Educational Psychology, 79, 347-362.

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

D'Agostino, J. V., Welsh, M. E., & Corson, M. E. (2007). Instructional sensitivity of a state's standards-based asssessment. Educational Assessment, 12, 1-22.

D'Andrade. (1992). Schemas and motivation. In R. G. D. A. C. Strauss (Ed.), Human motives and cultural models (pp. 23-44). Cambridge: Cabridge University Press.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. Seattle: Center for the Study of Teaching and Policy, University of Washington.

de Ribaupierre, A., & Rieben, L. (1995). Individual and situational variability in cognitive development. Educational Psycologist, 30(1), 5-14.

DeJong, D. H. (n.d.). Promises of the past: A history of Indian education in the United States. Golden, Colorado: North American Press.

Delpit, L. (1988). The silenced dialogue: Power and pedagogy in educating other people's children. Harvard Educational Review, 58, 280-298.

Department of Education. (2005). 2005 NAEP Science Results. Retrieved August 16, 2007, from http://www.nationsreportcard.gov/science_2005

Driscoll, M. P. (2000). Psychology of Learning for Instruction. Boston, MA.: Allyn and Bacon.

Driver, R. (1990). Assessing the progress of children's understanding in science: A developmental perspective. In G. Hein (Ed.), The assessment of trends on elementary school programs (pp. 204-216). Grand Forks: The North Dakota Study Group.

Driver, R., Asoko, H., Leach, J., Mortimer, E., & Scott, P. (1994). Constructing scientific knowledge in the classroom. Educational Researcher, 23(7), 5-12.

Embretson, S. E. (2004). Measuring human intelligence with artificial intelligence: Adaptive item generation. In R. J. Sternberg & J. Pretz (Eds.), Cognition and Intelligence. New York: Cambridge University Press.

Ercikan, K. (1998). Translation effects in international assessments. International Journal of Educational Research, 29(6), 543-553.

Erickson, G. L., & Erickson, L. J. (1984). Females and science achievement: Evidence, explanations, and implications. Science Education, 68, 63-89.

Espinosa, L. M. (2005). Curriculum and assessment considerations for young children from culturally, linguistically, and economically diverse backgrounds. Psychology in the Schools, 42(8), 837-853.

Feuer, M. J., Holland, P. W., Bertenthal, M. W., Hemphill, F. C., & Green, B. F. (1998). Equivalency and Linkage of Educational Tests. Washington, DC: National Academy Press.

Fields, A. (2005). Discovering Statistics Using SPSS. Beverly Hills: Sage Publications.

Fischer, G. H. (1997). Unidimensional linear logistic Rasch models: The nominal categories model. In W. J. v. d. L. R. K. Hambleton (Ed.), Handbook of modern item response theory (pp. 225-243). New York: Springer-Verlag.

Frisbie, D. A. (2005). Measurement 101: Some fundamentals revisited. Educational Measurement: Issues & Practice, 21-29.

Gao, X., Shavelson, R., & Baxter, G. P. (1994). Generalizability of Large-Scale Performance Assessments in Science: Promises and Problems. Applied Measurement in Education, 7(4), 323-342.

Garden, R. A., Lie, S., Robitaille, D. F., Angell, C., Martin, M. O., Mullis, I. V., et al. (2006). TIMSS advanced 2008 assessment frameworks. Retrieved November 12, 2007, from http://timss.bc.edu/timss_advanced/frameworks.html

Garner, R. (1990). When children and adults do not use learning strategies: Towards a theory of settings. Review of Educational Research, 60, 517-529.

Gay, G. (2000). Culturally Responsive Teaching: Theory, Research and Practice. New York: Teachers College Press.

Ginsburg-Block, M. D., & Fantuzzo, J. W. (1998). An evaluation of the relative effectiveness of NCTM standards-based interventions for low-achieving urban elementary students. Journal of Educational Psychology, 90(3), 560-569.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 18, 519-521.

Glaser, R. (1992). Expert knowledge and processes of thinking. In D. F. Halpern (Ed.), Enhancing thinking skills in the sciences and mathematics (pp. 63-75). Hillsdale, NJ: Lawrence Erlbaum Associates.

Glaser, R., & Baxter, G. P. (1999). Assessing active knowledge. Paper presented at the Annual Meeting for the Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. Pittsburg: Learning Research and Development Center University of Pittsburg.

Glaser, R., & Pellegrino, J. W. (1979). Cognitive correlates and components in the analysis of individual differences. Intelligence, 3(3), 187-215.

Glaser, R., Raghavan, K., & Baxter, G. P. (1992). Cognitive theory as the basis for design of innovative assessment: Design characteristics of science assessments (No. 349). Los Angeles: University of California, CRESST.

Gliner, J. A., & Morgan, G. A. (2000). Research Methods in Applied Settings. Mahwah: Lawrence Erlbaum Associates.

Globerson, T. (1989). What is the relationship between cognitive style and cognitive development? In T. G. T. Zelniker (Ed.), Cognitive style and cognitive development (Vol. 20, pp. 71-85). Norwood: Ablex.

Goldberg, G. L., & Roswell, B. S. (2001). Are multiple measures meaningful?: Lessons from a statewide performance assessment. Applied Measurement in Education, 14(2), 125-150.

Greenfield, P. M. (1998). Culture as process: Empirical methods for cultural psychology. In Y. H. P. J.W. Berry, and J. Pandey (Ed.), Handbook of cross-cultural psychology (Vol. 1, pp. 301-346). Needham Heights, MA: Allyn & Bacon.

Greeno, J. G., Pearson, P.D., and Schoefeld, A.H. (1996). Implications for NAEP of research on learning and cognition. Report of a study commissioned by the National Academy of Education. Stanford, CA: Panel on the NAEP Trial State Assessment, Conducted by the Institute for Research on Learning.

Grigg, W., Lauko, M., & Brockway, D. (2006). The Nation's Report Card: Science 2005 (No. NCES 2006-466). Washington, D. C.: Department of Education, National Center for Education Statistics: U.S. Government Printing Office.

Haertel, E. a. H., J. (2005). A historical perspective on validity arguments for accountability testing (No. CSE Report 654). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Hanna, G. (1986). Sex differences in the mathematics achievement of eighth graders in Ontario. Journal of Research in Mathematics Education, 17(3), 231-237.

Harris, R. J., Schoen, L. M., & Lee, D. J. (1986). Culture-based distortion in memory for stories. In J. L. Armagost (Ed.), Papers from the 1985 Mid-America Linguistics Conference. Manhattan: Kansas State University Department of Speech.

Hatano, G., & Oura, Y. (2003). Reconceptualizing school learning using insight from expertise research. Educational Researcher, 32(8), 26-29.

Hattie, J., & Jaeger, R. M. (1998). Assessment and classroom learning: A deductive approach. Assessment in Education, 5, 111-122.

Hattie, J., Jaeger, R. M., & Bond, L. (1999). Persistent Methodological Questions in Educational Testing. Review of Research in Education, 24, 393-446.

Haydel, A. M. (2003). Using cognitive analysis to understand motivational and situational influences in science achievement. Paper presented at The Annual Meeting of the American Educational Research Association, Chicago, Il.

Henderson, A. T., & Berla, N. (1994). A new generation of evidence: The family is critical to student achievement (No. ED3750968). Washington, DC: Center for Law and Education.

Herman, J. L. (2004). The effects of testing on instruction. In Redesigning Accountability Systems for Education (pp. 141-166). New York: Teachers College Press.

Hidalgo, N. M., Siu, S., Bright, J. A., Swap, S. M., & Epstein, J. L. (1995). Research on families, schools, and communities: A multicultural perspective. In J. Banks & C. A. A. and Banks (Eds.), Handbook of Research on Multicultural Education (pp. 498-524). New York: Simon & Schuster Macmillan.

Hotano, G., & Oura, Y. (2003). Commentary: Reconceptualizing school learning using insight from expertise research. Educational Researcher, 32(8), 26-29.

Johnson, S. (1987). Gender differences in science: Parallels in interest, experience, and performance. International Journal of Science Education, 9, 467-481.

Jones, L. R., Mullis, I. V., Raizen, S. A., Weiss, I. R., & Weston, E. A. (1992). The 1990 science report card NAEP assessment of fourth, eighth, and twelfth graders. Washington, DC: National Center for Educational Statistics.

Jovanovic, J., Solano-Flores, G., & Shavelson, R. (1994). Performance-based assessments: Will gender differences in science achievement be eliminated? Education and Urban Society, 26(3), 352-366.

Kane, T. J., & Staiger, D. O. (2001). Improving school accountability measures. Retrieved April 5, 2006, from http://www.dartmouth.edu/~dstaiger/Papers/kanestaigerbrookings.pdf

Kemler Nelson, D. G., & Smith, J. D. (1989). Holistic and analytic processing in reflection-impulsivity in cognitive development. In T. G. T. Zelniker (Ed.), Cognitive Style and Cognitive Development (pp. 116-140).

Kennedy, J. J. (1970). The eta coefficient in complex ANOVA designs. Educational and Psychological Measurement, 30, 885-890.

Kenney, P. A., & Lindquist, M. M. (2000). Students' performance on thematically related NAEP tasks. In Results from the 7th Mathematics Assessment of the National Assessment of Educational Progress. Reston, VA.: National Council of Teachers of Mathematics.

Kher, N., Schmidt, W., Houang, R., & Zou, Z. (2007). High school mathematics trajectories: Connecting opportunities to learn with student performance. Paper presented at the Annual Meeting of the American Education Research Association, San Fransisco, CA.

Klein, S. P., Jovanovic, J., Stecher, B.M., McCaffrey, D., Shavelson, R.J., Haertel, E., Solano-Flores, G., and Comfort, K. (1997). Gender and racial/ethnic differences on performance assessments in science. Educational Evaluation and Policy Analysis, 19(2), 83-97.

Koelsch, N. (1994). Chinle Public Schools portfolio assessment implementation manual. San Francisco: Far West Laboratory.

Koelsch, N., & Trumbell, E. (1996). Portfolios: Bridging Cultural and Linguistic Worlds. In Writing Portfolios in the Classroom (pp. 261-285). New Jersey: Lawrence Erlbaum Associates Publishers.

Kupermintz, H., Le, V., & Snow, R. (1999). Construct validation for mathematics achievement: Evidence from interview procedures (No. CSE Technical Report 493). Los Angeles, Center for the Study of Evaluation: University of California.

Lambert, N. M., & McCombs, B. L. (1998). How students learn: Reforming schools through learner-centered education. Washington, DC: American Psychological Association.

Lane, S., Wang, N., & Magone, M. (1996). Gender related differential item functioning on a middle school mathematics performance assessment. Educational Measurement: Issues & Practice, 15(4), 21-27.

Langer, J. A., Applebee, A. N., Mullis, V. S., & Foertsch, M. A. (1990). Learning to read in our nation's schools: Instruction and achievement at grades 4, 8, and 12. Princeton: Educational Testing Service.

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. Science Edition, 208, 1335-1342.

Lave, J., & Wenger, E. (1991). Situated Learning: Legitimate Peripheral Participation: Cambridge University Press.

Lawrenz, D., Huffman, F., & Welch, W. (2001). The science achievement of various subgroups on alternative assessment formats. Science Education, 85, 279-290.

Leinhardt, G., & Seewald, A. (1981). Overlap: What's tested, what's taught? Journal of Educational Measurement, 18, 85-96.

Linn, M. C., de Benedictis, T., Delucchi, K., Harris, A. & Stage, E. (1987). Gender differences in national assessment of educational progress science items: What does i don't know really mean? Journal of Research in Science Teaching, 24(3), 267-278.

Linn, M. C. (1992). Gender differences in educational achievement. Paper presented at the 1991 Inivitational Conference of the Educational Testing Service, Princeton, N. J.

Lockhart, R. S., & Craik, F. M. (1990). Levels of processing: A retrospective commentary on a framework for memory research. Journal of Canadian Psychology, 44(1), 87-112.

Loevinger, J. A. (1947). A systematic approach to the construction and evaluation of tests of ability. Psychological Monographs, 61(4).

Loevinger, J. A. (1965). Person and population as psychometric concepts. Psychological Review, 72(2), 143-155.

Lohman, D. F. (1994). Component scores as residual variation (or why the intercept correlates best). Intelligence, 19, 1-11.

Lohman, D. F. (2000). Complex information processing and intelligence. In R. J. Sternberg (Ed.), Handbook of Intelligence (pp. 285-340). Cambridge: Cambridge University Press.

Lopez, A., Atran, S., Coley, J. D., & Medin, D. L. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. Cognitive Psychology, 32(3), 251-295.

Lynch, E., & Hanson, M. (2004). Developing cross-cultural competence: A guide for working with children and their families. Baltimore, MD:: Paul H. Brookes.

Madaus, G. F., Airasian, P. W. & Kellaghan, T. (1980). School effectiveness: A reassessment of the evidence. New York: McGraw-Hill.

Madaus, G. F., Stufflebeam, D., & Scriven, M. S. (1983). Program evaluation: A historical overview. In Evaluation Models (Vol. 49, pp. 3-22): Springer Netherlands.

Mager, R. F. (1962). Preparing instructional objectives. Belmont, CA: Fearon Publishers.

Margolis, H. (1990). Patterns, Thinking, and Cognition: A Theory of Judgement. Chicago: University of Chicago Press.

Marshall, S. P. (1995). Schemas in problem solving. Cambridge: Cambridge University Press.

Matlin, M. W. (2005). Cognition. Hoboken, NJ: John Wiley & Sons, Inc.

Mayer, R. E. (1998). Cognitive theory for education: What teachers need to know. In N. M. B. L. M. Lambert (Ed.), How Students Learn (pp. 353-377). Washington, DC: American Psychological Association.

Mazzeo, J., Schmitt, A., & Bleistein, C. (1991). Sex-related differences on constructed-response and multiple-choice sections of advanced placement examinations: Three exploratory studies. Princeton: Educational Testing Service.

McCall, W. A. (1922). How to measure in education. New York: MacMillan.

McDonnell, L. M., McLaughlin, M. J., & Morison, P. (Eds.). (1997). Educating one and all: Students with disabilities and standards-based reform. Washington D.C.: National Academy of Education.

McGill-Franzen, A., & Allington, R. L. (1993). Flunk 'em or get them reclassified: The contamination of primary grade accountability data. Educational Researcher, 22(1), 19-22.

Meisels, S. J. A.-B., S.; Xue, Y.; Bickel, D.; Son, S.; Nicholson, J. (2003). Creating a system of accountability: The impact of instructional assessment on elementary children's achievement test scores. Retrieved 10/24/2006, from http://epaa.asu.edu/epaa/v11n9/

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement. New York: American Council on Education and MacMillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23(2), 13-23.

Minsky, M. (1985). The Society of Mind. New York: Simon and Schuster.

Minstrell, J. (2000). Student thinking and related assessment: Creating a facet-based learning environment. In J. W. P. N. S. Raju, M. W. Bertenthal, K. J. Mitchell, & L. R. Jones (Ed.), Grading the nation's report card: Research from the evaluation of NAEP (pp. 44-73). Washington, D.C.: National Academies Press.

Mislevy, R. (1996). Test theory reconceived. Journal of Educational Measurement, 33, 379-416.

Montague, M. (1997). Cognitive strategy instruction in mathematics for students with learning disabilities. Journal of Learning Disabilities, 30, 164-177.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. Psychological Review, 92, 289-316.

National Academy of Sciences. (1996). National Science Education Standards. Washington, DC: National Academy Press.

National Research Council. (2001). Knowing what students know: The science and design of educational assessment. Washington , DC: National Academy Press.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Pardos, Z. A., Heffernan, N. T., Anderson, B., & Heffernan, C. L. (2006). Using fine-grained skill models to fit student performance with Bayesian Networks. Paper presented at the 8th International Conference on Intelligent Tutoring Systems, Jhongli, Taiwan.

Peak, L. (1996). Pursuing excellence (No. NCES 97-198). Washington, DC: US Government Printing Office.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. Review of Research in Education, 24, 307-353.

Peske, H. G., & Haycock, K. (2006). Teaching inequality: How poor and minority students are shortchanged on teacher quality. Washington, D.C.: The Education Trust.

Peverly, S. T. (1991). Problems with the knowledge-based explanation of memory and development. Review of Educational Research, 61, 71-93.

Phillips, S. E. (1993). Legal Implications of High-Stakes Assessment: What States Should Know. Michigan State University, North Central Regional Educational Laboratory, Illinois.

Piaget, J., & Inhelder, B. (1956). The Child's Conception of Space. London: Routledge.

Pintrich, P. R., Marx, R. W., & Boyle, R. A. (1993). Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. Review of Educational Research, 63(2), 167-199.

Randolf, J. R., & Edmondson, R. S. (2005). Using the binomial effect size display (BESD) to present the magnitude of effect sizes to the evaluation audience. Retrieved November 20, 2007, from www.unc.edu/~chambles/psyc256/BESD.html

Rasch, G. U. (1961). On general laws and meaning of measurement in psychology. Paper presented at the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, California.

Resnick, L. B. (1979). The future of IQ testing. In R. J. Sternberg & D. K. Detterman (Eds.), Human Intelligence: Perspectives on its Theory and Measurement (pp. 203-215). Norwood, NJ: Ablex.

Reuchlin, M. (1978). Vicarious processes and individual differences. Journal of Psychology, 2, 133-145.

Rodriguez, A. J. (1997). The dangerous discourse of invisibility: A critique of the National Research Council's national science education standards. Journal of Research in Science Teaching, 34, 19-37.

Rodriguez, M. C. (2004). The role of classroom assessment in student performance on TIMMS. Applied Measurement in Education, 17(1), 1-24.

Rosa, M. (2007). How districts shortchange low-income and minority students: Funding gaps 2006. Washington, D.C.: The Education Trust.

Rowley, G. L. (1974). Which examinees are most favoured by the use of multiple choice tests? Journal of Educational Measurement, 11(1), 15-23.

Royer, J. M., Cisero, C. A., & Carlo, M. S. (1993). Techniques and procedures for measuring cognitive skills. Review of Educational Research, 63, 201-243.

Ruiz-Primo, M. A., Shavelson, R., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. Journal of Research in Science Teaching, 39(5), 369-393.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. Spiro, B. C. Bruce & W. F. Brewer (Eds.), Theoretical Issues in Reading Comprehension (pp. 33-58). Hillsdale, NJ: Erlbaum.

Sanchez, M., & Brisk, M. (2004). Teachers assessment practices and understandings in a bilingual program. NABE Journal of Research and Practice, 2(1), 193-208.

Schmidt, W., McKnight, C. C., & Raizen, S. A. (1997). A Splintered Vision. Boston: Kluwer Academic.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. Psychological Methods, 1, 115-129.

Sencar, S., & Eryilmaz, A. (2004). Factors mediating the effect of gender on ninth-grade Turkish students' misconceptions concerning electric circuits. Journal of Research in Science Teaching, 41(6), 603-616.

Shaftel, J., Belton-Kocher, E., Glasnapp, D. & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. Educational Assessment, 11(2), 105-126.

Shallert, D. L. (1976). Improving memory for prose: The relationship between depth of processing and context. Journal of Verbal Learning and Verbal Behavior, 15, 621-632.

Shannon, G. S., & Bylsma, P. (2007). Nine characteristics of high-performing schools. Olympia: Office of Superintendent of Public Instruction.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessment: Political rhetoric and measurement reality. Educational Researcher, 21(4), 22-27.

Shepard, L., Taylor, G., & Betebenner, D. (1998). Inclusion of limited-english-proficient students in Rhode Island's Grade 4 mathematics performance assessment (No. CSE Technical Report 486). Santa Cruz: Center for Research on Education, Diversity and Excellence.

Shulman, L. S., & Quinlan, K. M. (1996). The Comparative Psychology of School Subjects. In R. C. Calfee (Ed.), Handbook of educational psychology (pp. 399-422). New York: Prentice Hall International.

Sicoly, F. (2002). Stability of school-level scores from large-scale student assessments. Applied measurement in education, 15, 173-185.

Siegler, R. S. (1976). Three aspects of cognitive development. Cognitive Psychology, 8, 481-520.

Silver, E. A., Cengiz, A., & Stylianou, D. A. (2000). Students' performance on extended constructed-response tasks. In Results from the 7th Mathematics Assessment of the National Assessment of Educational Progress (pp. 301-341). Reston, VA.: National Council of Teachers of Mathematics.

Sinclair, A. S. (1994). Prediction making as an instructional strategy: Implications of teacher effects on learning, attitude toward science, and classroom participation. Journal of Research and Development in Education, 27, 153-161.

Skinner, B. F. (1938). The Behavior of Organisms. New York: Appleton-Century-Crofts.

Smith, E. R., & Tyler, R. W., & the Evaluation Staff. (1942). Appraising and recording student progress - The Adventure in American Education Series (Vol. III). New York: Harper & Bros.

Smith, M. U. (1991). Toward a unified theory of problem solving: Views from the content domain. Hillsdale, NJ: Lawrence Erlbaum Associates.

Smith, M. U., & Good, R. (1984). Problem solving and classical genetics: Successful versus unsuccessful performance. Journal of Research in Science Teaching, 21, 895-912.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. Linn (Ed.), Educational Measurement (pp. 263-331). New York: MacMillan.

Solano-Flores, G. (2002, April 1-5, 2002). Cultural Validity:The Need for a Socio-Cultural Perspective in Educational Measurement. Paper presented at the American Education Research Association, New Orleans, LA.

Solano-Flores, G., & Nelson-Barber, S. (2000). Cultural Validity of Assessments and Assessment Development Procedures. Paper presented at the Annual Meeting of the American Education Research Association, New Orleans, LA.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. Journal of Research in Science Teaching, 38(5), 553-573.

Solano-Flores, G., & Shavelson, R. J. (1997). Development of performance assessments in science: Conceptual, practical and logistical issues. Educational Measurement: Issues & Practice, 16(3), 16-25.

Solano-Flores, G., & Trumbell, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. Educational Researcher, 32(2), 3-13.

Stecher, B., Klein, S., Solano-Flores, G., McCaffery, D., Robyn, A., Shavelson, R., et al. (2000). The effects of content, format, and inquiry level on science performance assessment scores. Applied Measurement in Education, 13(2), 139-160.

Steele, C. M., & Aronson, J. (1995). Stereotype type threat and the intellectual test performance of African Americans. Journal of Personality and Social Psychology, 69(5), 797-811.

Sugrue, B. (1997). Specifications for the Design of Problem-Solving Assessments in Science (No. 387). Los Angeles: University of California, CRESST.

Swisher, K., & Deyhle, D. (1987). Styles of learning and learning of styles: Education conflicts for American Indian/Alaska Native youth. Journal of Multilingual and Multicultural Development, 8(4), 345-360.

Tabachnick, B. G., & Fidell, L. S. (2007). Using Multivariate Statistics. Boston: Allyn & Bacon.

Terman, L. M. (1916). The measurement of intelligence. Boston: Houghton Mifflin.

Tharp, R. (1982). The effective instruction of comprehension: Results and description of the Kamehameha Early Education Program. Reading Research Quarterly, 71(4), 503-527.

Thomson, D. M., & Tulving, E. (1970). Associative encoding and retrieval: Weak and strong cues. Journal of Experimental Psychology, 86, 255-262.

Thorndike, E. L. (1918). The nature, purposes, and general methods of measurements of educational products. Measurement of educational products, 16-24.

Thorndike, E. L. (1926). The Original Nature of Man (Vol. 1). New York: Teachers College.

Thorndike, E. L. (1931). All the world's a chess-board. Speculum, 6(3), 461-465.

Thurstone, L. L. (1927). A law of comparative judgment. Psychological Review, 101(2), 266-270.

TIMSS. (2004). Trends in international mathematics and science study. Retrieved November 7, 2007, from timss.bc.edu

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. Psychological Review, 80, 352-373.

Tyler, R. W. (1948). Educability and the schools. The Elementary School Journal, 49(4), 200-212.

Van de Vijver, J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross cultural assessment. European Journal of Psychological Assessment, 13(1), 29-37.

Van-Largen, A., Bosker, R., & Dekkers, H. (2006). Exploring cross-national differences in gender gaps in education. Educational Research and Evaluation, 12(2), 155-177.

Von Secker, C. E., & Lissitz, R. W. (1999). Estimating the impact of instructional practices on student achievement in science. Journal of Research in Science Teaching, 36(10), 1110-1126.

Vosniadou, S. (1994). Universal and culture-specific properties of children's mental models of the earth. In Mapping the mind: Domain specificity in cognition and culture (pp. 412-430). Cambridge: Cambridge University Press.

Webb, N. L. (1997). Criteria for alignment of expectations and assessments in mathematics and science education. Washington, DC: National Institute for Science Education and Council of Chief State School Officers Research Monograph No. 6.

White, B. Y., & Frederiksen, J. Y. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. Cognition and Instruction, 16(1), 3-118.

Willhoft, J., & Lee, Y. (2005, December 7-9, 2005). Introduction to Item Response Theory (IRT). Paper presented at the 21st Annual Washington State Assessment Conference, Seattle, WA.

Williamson, D. M., Bauer, M., Steinberg, L. S., Mislevy, R. J., Behrens, J. T., & DeMark, S. F. (2004). Design rationale for a complex performance assessment. International Journal of Testing, 4(4), 303-332.

Winfield, L. F. (1993). Investigating test content and curriculum content overlap to assess opportunity to learn. Journal of Negro Education, 62(3), 288-310.

Wittgenstein, L. (1963). The Philosophical Investigations. New York: The Macmillan Company.

Woltz, D. J. (2003). Implicit cognitive processes as aptitudes for learning. Educational Psycologist, 38(2), 95-104.

Wright, B. D., & Stone, M. H. (1979). Best Test Design: Rasch Measurement. Chicago: Mesa Press.

Yoon, B. R., L. B. (1998). Instructional validity, opportunity to learn and equity: New standards examinations for the California Mathematics Renaissance. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2004). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. Educational Assessment, 9, 61-78.

Zumback, J., & Reimann, P. (2002). Assessment of a goal-based scenario approach: A hypermedia comparison. Instructional Science, 30, 243-267.

Zurawski, C. (2004). Teachers matter: Evidence from value-added assessments. Washington D.C.: American Education Research Association.

APPENDICES

APPENDIX A

AN EXAMPLE OF A MODEL OF COGNITION

Appendix A

An Example of a Model of Cognition

Understanding Gravity in Fluid Mediums (Minstrell, 2000, p.52)

Cluster Number	Description of Understanding
310 (correct)	**Pushes from above and below by a surrounding fluid medium lend a slight support (net upward push due to differences in depth pressure gradient).**
310-1 (incorrect)	The difference between the upward and downward pushes by the surrounding air results in a slight upward support or buoyancy.
310-2 (incorrect)	Pushes from above and below an object in a liquid medium yield a buoyant upward force due to the larger pressure from below.
315 (correct)	**Surrounding fluids exert equal pushes all around an object.**
315-1 (incorrect)	Air pressure has no up or down influence.
315-2 (incorrect)	Liquid presses equally from all sides regardless of depth.

APPENDIX B

WASHINGTON ASSESSMENT OF STUDENT LEARNING

Appendix B

WASL – Washington Assessment of Student Learning

A Component of the Washington State Assessment Program

Using Results to Improve

Student Learning

Science
Grade 8

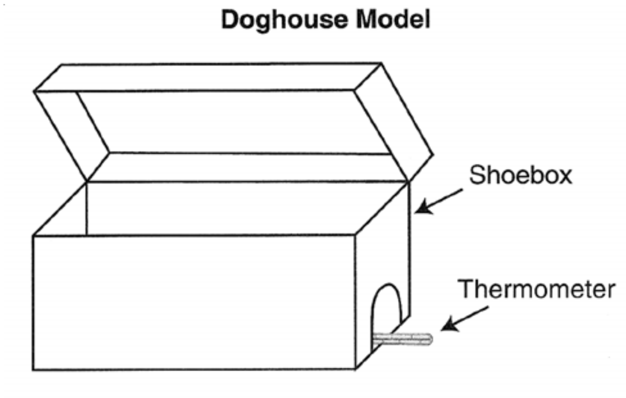
2005 Released Scenarios and Items

Published by the Science Assessment Team of the Washington Office of the Superintendent of Public Instruction on September 28, 2005.

Scenario Question "In the Doghouse"

Directions: Use the following information to answer numbers 1 through 5.

Paul and Dalia live in eastern Washington. They want to build a new doghouse for their dog, Fido, that will keep him warm in the winter. They built a model of their doghouse using a shoebox as shown in the diagram below.



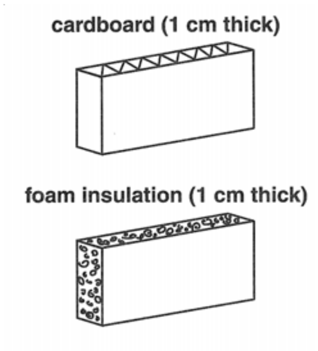
Paul and Dalia conducted the following investigation using their doghouse model.

Question:
How does insulating the walls and ceiling of a doghouse model with different materials affect the inside temperature of the doghouse model?

Hypothesis (prediction):
The inside temperature of the doghouse model will be warmest when insulated with foam insulation. The reason for our prediction is foam insulation is used when building houses for people.

Materials:
doghouse model
timer
thermometer
freezer
insulating materials

- cardboard
- foam insulation



Procedure:

Measure the inside temperature of the doghouse model and record as “Before” temperature.
Place the doghouse model in the freezer.
After five minutes in the freezer, measure the inside temperature of the doghouse model and record as “After” temperature.
Remove the doghouse model from the freezer and let the model return to room temperature.
Insulate the inside walls and ceiling of the doghouse model with cardboard insulation, then repeat steps 2, 3, and 4.
Repeat step 5 using foam insulation.
Starting at Step 1, repeat the entire investigation twice for Trials 2 and 3.

Data:

Insulation vs. Inside Temperature

Insulation	**Inside Temperature (˚C)**					
	Before Trials			**After Trials**		
	1	2	3	1	2	3
None	22	21	21	4	3	4
Cardboard	22	21	22	13	12	12
Foam	21	22	22	11	11	10

In the Doghouse

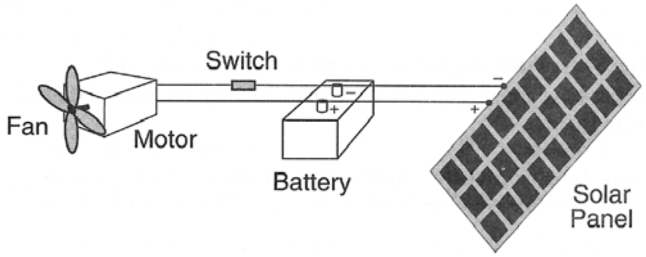
1 Which variable in the investigation was the responding (dependent) variable?

- o **A**. Temperature inside the freezer
- o **B**. Temperature inside the doghouse model
- o **C**. Length of time the doghouse model was in the freezer
- o **D**. Type of insulating material added to the doghouse model

2 Sunlight can raise the inside temperature of a real doghouse. Which of the following changes to a real doghouse would result in warmer inside temperatures due to sunlight?

- o **A**. Covering the doghouse roof with a shiny metal material
- o **B**. Making the doghouse roof thicker with the same material
- o **C**. Changing the doghouse roof from a light color to a dark color
- o **D**. Adding a small metal air vent to the middle of the doghouse roof

3 Paul and Dalia considered using a solar-powered fan in Fido's doghouse. The solar fan system is pictured below.



Where is energy stored in the solar fan system?

- o **A**. Solar panel
- o **B**. Battery
- o **C**. Motor
- o **D**. Fan

4 The entrance to the doghouse is large. Dalia thinks this opening will have an effect on the inside temperature of the doghouse. How could Dalia change the doghouse entrance to keep more heat energy inside?

Be sure to:

- Describe how the doghouse entrance could be changed.
- Explain how the changes would help keep more heat energy inside.

Use words, labeled pictures, and/or labeled diagrams in your response.

In the Doghouse

Scoring rubric for item number 4:

Evaluating Potential Solutions

A **2-point response** demonstrates that the student understands the GLE: DE03 Analyze multiple solutions to a problem or challenge.

The student describes how the doghouse entrance could be changed.
AND
The student explains how the changes would help keep more heat energy inside.

Examples:

How the entrance could be changed	How the change would keep more energy inside
• *Smaller opening* • *Flap or cover over the opening*	• *Warm air would stay inside* • *Cold air could not come into the doghouse* • *Wind could not blow into the doghouse*

A **1-point response** demonstrates that the student has partial understanding of the GLE.

The student describes how the doghouse entrance could be changed **but** does not explain, or only provides a partial explanation, as to how the changes would help keep more heat energy inside.

A **0-point response** demonstrates that the student shows little or no understanding of the GLE.

Notes:

1. In the discussion of the movement of heat or cold, the carrier of the heat or cold needs to be identified (e.g. air, wind, etc).

In the Doghouse

Annotated example of a 2-point response:

4 The entrance to the doghouse is large. Dalia thinks this opening will have an effect on the inside temperature of the doghouse. How could Dalia change the doghouse entrance to keep more heat energy inside?

Be sure to:

- Describe how the doghouse entrance could be changed.
- Explain how the changes would help keep more heat energy inside.

Use words, labeled pictures, and/or labeled diagrams in your response.

seal door is closed
no cold air can go in.
Then shuts, sealing the warm air inside

Dalia could build a swinging door to the doghouse. So, when the dog is inside the house, the heat

is kept in. Also, when the dog goes out the swinging door will shut behind Fido and seal the door

so that no cold air will go in.

Annotation	Score Point
cribes the change to the entrance: *…build a swinging door…*	1
Describes how the change would help keep more heat energy inside: *…when the dog is inside the house, the heat is kept in… and seal the door so that no cold air will go in.*	1
Total Score Points	**2**

In the Doghouse

Annotated example of a 1-point response:

4 The entrance to the doghouse is large. Dalia thinks this opening will have an effect on the inside temperature of the doghouse. How could Dalia change the doghouse entrance to keep more heat energy inside?

Be sure to:

- Describe how the doghouse entrance could be changed.
- Explain how the changes would help keep more heat energy inside.

Use words, labeled pictures, and/or labeled diagrams in your response.

She could use a dark covering for the door like a dark flap.

Annotation	Score Point
cribes the change to the entrance: *She could use a dark covering for the door like a dark flap.*	1
Describes how the change would help keep more heat energy inside: None	0
Total Score Points	**1**

In the Doghouse

Annotated example of a 0-point response:

4 The entrance to the doghouse is large. Dalia thinks this opening will have an effect on the inside temperature of the doghouse. How could Dalia change the doghouse entrance to keep more heat energy inside?

Be sure to:

- Describe how the doghouse entrance could be changed.
- Explain how the changes would help keep more heat energy inside.

Use words, labeled pictures, and/or labeled diagrams in your response.

Put a thicker layer on inside or outside the dog house out would be easier

Annotation	Score Point
cribes the change to the entrance: *Put a thicker layer on inside or outside the dog house out would be easier*. Not a change to the entrance. Discusses insulation.	**0**
Describes how the change would help keep more heat energy inside: None	**0**
Total Score Points	**0**

In the Doghouse

5 Dalia wondered if a large, well-insulated doghouse would stay warmer in winter than a small one. She asked, “How does the size of a well-insulated doghouse model affect the inside temperature of the doghouse model?” Plan an investigation to answer Dalia’s question.

Be sure to include:

- Hypothesis (prediction) of the investigation results
- Procedure that includes:
    - logical steps to do the investigation
    - one controlled (kept the same) variable
    - one manipulated (changed) variable
    - one responding (dependent) variable
    - how often measurements should be taken and recorded.

Question: How does the size of a well-insulated doghouse model affect the inside temperature of the doghouse model?

Hypothesis (prediction):

Materials:

You may use the space below for a labeled diagram to support your procedure.

Procedure:

In the Doghouse

Scoring rubric for item number 5:

Performance Description	**Value Points**
A **4-point response** demonstrates that the student has understanding of the GLE: IN02 Understand how to plan and conduct scientific investigations.	**8-9**
A **3-point response** demonstrates that the student partially understands the GLE.	**6-7**
A **2-point response** demonstrates that the student has limited understanding of the GLE.	**4-5**
A **1-point response** demonstrates that the student has very little understanding of the GLE.	**2-3**
A **0-point response** demonstrates that the student has almost no understanding of the GLE.	**0-1**

Attributes of a Controlled Investigation for Awarding Value Points

Investigation Attributes	**Description of Attribute**	**Value Point**
Prediction	The prediction portion of the hypothesis must answer the given question including the effect of the manipulated (changed) variable *(the size of doghouse model)* on the responding (dependent) variable (*inside temperature*). *The smaller the doghouse model, the more constant the inside temperature…*	1
Prediction Reason	A hypothesis must give a related reason for the prediction. …*because there is less air to keep warm.*	1
Materials	A list of the minimum materials needed to perform the procedure: • *well-insulated shoebox (or similar box) either larger or smaller than the original box used, or at least two different size boxes (if two boxes were used in this investigation);* • *timer;* • *thermometer*; • *freezer or similar means to manipulate temperature.* Attribute Notes: 1. The 'right' amount of ingredients (e.g. 'x' mL or 'y' grams) needed to carry out the procedure do not need to be given in the material	1

Attributes of a Controlled Investigation for Awarding Value Points (continued)

Investigation Attributes	**Description of Attribute**	**Value Point**
Procedure:	The written or diagrammed procedure is evaluated as follows.	Up to 6
Controlled (kept the same) Variable	At least one controlled (kept the same) variable must be identified or implied in the procedure or the materials list *(e.g. the same insulation, time in the freezer).*	1
Manipulated (changed) Variable	Only one manipulated (changed) variable *(at least two sizes of well-insulated doghouse models)* is identified or implied in the procedure or data table (if given).	1
Responding (dependent) Variable	The responding (dependent) variable *(inside temperature of the doghouse)* is identified or implied in the procedure or data table (if given).	1
Record Measurements	The procedure states or implies measurements are recorded periodically or gives a data table. Attribute Note: 1. If artificial data for the responding variable is given, no value point may be awarded. 2. The phrase 'take measurement' cannot be used to mean record.	1
Trials are Repeated	More than one trial for all conditions is planned, or implied in a data table, to measure the responding (dependent) variable.	1
Logical Steps	The steps of the procedure are detailed enough to repeat the procedure effectively (examples of illogical steps: no ending time indicated, states "Set up as diagrammed" but diagram is inadequate, recording vague data or results).	1
	Total Value Points Possible	**9**

	list. 2. If pre-measured amounts of materials are listed in the materials list, a measuring device may not be needed in the materials list. 3. Standard Classroom Materials do not need to be listed: paper, pencil, safety equipment (e.g. goggles, aprons, gloves, tongs).	

Attributes of a Controlled Investigation for Awarding Value Points (continued)

Notes:

1. If the response does not plan an appropriate procedure for the given question, the response may not earn any of the possible procedure value points.
   Examples:
   a) repeats the procedure from the scenario
   b) plans for only one measurement to be taken thus there are no controlled or manipulated variables
   c) purposefully changes more than one variable simultaneously
   d) give a procedure that is too vague to possibly be appropriate
   e) gives a prediction instead of a procedure
2. If the response names a bulleted attribute listed after "Procedure that includes:" without including that attribute in the procedure, the attribute point may be not be credited. When a bulleted attribute is named and implied in the response, both must be correct to credit.
3. Vagueness in procedural steps shall be clarified as follows:
   a) Vague materials used in the procedure (e.g. 1mL) may be credited if the vagueness is clarified in the materials list (e.g. 1mL of insecticide)
   b) Measuring a vague parameter (e.g. size of plant instead of height) may be credited as a responding variable but is too vague to repeat, so no point can be credited for 'logical steps.'
   c) The term "repeat" at the end of a step refers to that step only.
   d) The term "repeat" as a separate step (or in a new paragraph) refers to the whole procedure.
   e) The term "repeat" when qualified as "if necessary" cannot be credited.
   f) A vague action that calls for the manipulated variable to be changed (e.g. increase the temperature by 5˚ C) without indicating how many times, give no end to the investigation so logical steps cannot be credited.
   g) A vague action that calls for the manipulated variable to be changed (e.g. increase the temperature by 5˚ C) without indicating how many times, cannot be credited for more than two conditions.

Scoring rubric for item number 5, continued:

In the Doghouse

Annotated example of a 4-point response:

5 Dalia wondered if a large, well-insulated doghouse would stay warmer in winter than a small one. She asked, “How does the size of a well-insulated doghouse model affect the inside temperature of the doghouse model?” Plan an investigation to answer Dalia’s question.

Be sure to include:

- Hypothesis (prediction) of the investigation results
- Procedure that includes:
  - logical steps to do the investigation
  - one controlled (kept the same) variable
  - one manipulated (changed) variable
  - one responding (dependent) variable
  - how often measurements should be taken and recorded.

Question: How does the size of a well-insulated doghouse model affect the inside temperature of the doghouse model?

Hypothesis (prediction): *the inside temperature of a large, well-insulated doghouse would stay*

warmer in the winter than the small one. the reason for my prediction is because the larger

Doghouse has larger walls, therefore having more insulation to trap the heat.

Materials: *large doghouse model (1), small doghouse model (1), timer (1), freezer,*

thermometer (1), foam insulation

Annotated example of a 4-point response (continued):

large well-insulated doghouse model vs. Small.

	temperature (°c)					
model	Before trials			After trials		
size	1	2	3	1	2	3
Small						
large						

Procedure: *1. Insulate the walls and ceiling of the small doghouse model with foam insulation. Repeat with large doghouse model. 2. Measure the inside temperature of the small and large doghouse model and record as the "Before" temperature. 3. Place the doghouse models into the freezer. 4. Use timer to time, after 5 minutes in the freezer, measure the inside temperature of both doghouse models and record as "After" temperature. 5. Starting with step 1, repeat the entire investigation two more times, recording the results each time. 6. Clean up.*

In the Doghouse

Annotation of the 4-point response:

Investigation Attributes	**Value Point**	**Annotation**
Prediction	1	*the inside temperature of a large… doghouse would stay warmer… than the small one.*
Prediction Reason	1	*…because the larger doghouse has larger walls, therefore having more insulation to trap the heat.*
Materials	1	Minimum materials are listed.
Controlled (kept the same) Variable	1	Same freezer
Changed (manipulated) Variable	1	Small and large doghouse models
Measured (responding) Variable	1	*Measure the inside temperature*
Record Measurements	1	*Record as 'after' temperature.*
Trials are Repeated	1	Three trials: *Repeat the entire investigation two more times*
Logical Steps	1	Can be effectively repeated.
Total	**9**	**4** Score Points

In the Doghouse

Annotated example of a 3-point response:

5 Dalia wondered if a large, well-insulated doghouse would stay warmer in winter than a small one. She asked, “How does the size of a well-insulated doghouse model affect the inside temperature of the doghouse model?” Plan an investigation to answer Dalia’s question.

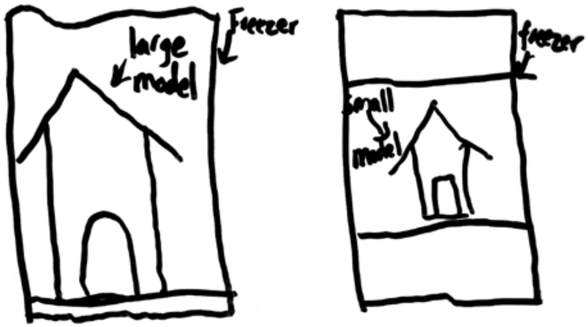
Be sure to include:

- Hypothesis (prediction) of the investigation results
- Procedure that includes:
  - logical steps to do the investigation
  - one controlled (kept the same) variable
  - one manipulated (changed) variable
  - one responding (dependent) variable
  - how often measurements should be taken and recorded.

Question: How does the size of a well-insulated doghouse model affect the inside temperature of the doghouse model?

Hypothesis (prediction): *I think the less space in a doghouse the warmer it will stay.*

Materials: *Small doghouse model, large doghouse model, timer, thermometer in ˚f, freezer,*



Procedure: *Build both doghouse models constructed of the same material, record temperature before placing in freezer, place in freezer for 20 minutes, record temature after taking out of*

freezer. Repeat 3 times. – controlled variable – temp in freezer - manipulated variable – The

temp before being placed in freezer. – responding variable – Temp of doghouse after taken out of freezer.

In the Doghouse

Annotation of the 3-point response:

Investigation Attributes	**Value Point**	**Annotation**
Prediction	1	*…the less space in a doghouse the warmer it* (doghouse) *will stay*
Prediction Reason	0	No reason stated
Materials	1	Minimum materials are listed.
Controlled (kept the same) Variable	1	Identified as *temp in freezer*
Changed (manipulated) Variable	0	Identified incorrectly as *the temp before being placed in freezer*
Measured (responding) Variable	1	Identified as *temp of doghouse after taken out of freezer*
Record Measurements	1	*Record temperature after taking out of freezer*
Trials are Repeated	1	Four trials: *Repeat 3 times*
Logical Steps	1	Can be effectively repeated.
Total	**7**	**3** Score Points

In the Doghouse

Annotated example of a 2-point response:

5 Dalia wondered if a large, well-insulated doghouse would stay warmer in winter than a small one. She asked, “How does the size of a well-insulated doghouse model affect the inside temperature of the doghouse model?” Plan an investigation to answer Dalia’s question.

Be sure to include:

- Hypothesis (prediction) of the investigation results
- Procedure that includes:
    - logical steps to do the investigation
    - one controlled (kept the same) variable
    - one manipulated (changed) variable
    - one responding (dependent) variable
    - how often measurements should be taken and recorded.

Question: How does the size of a well-insulated doghouse model affect the inside temperature of the doghouse model?
Hypothesis (prediction): *I think it would effect the dog house.*
Materials: *1 dog house (small), 1 thermometer, 1 dog house (large), 1 lamp, ice cubes (about 10)*
Procedure: *1.) with the small dog house, place a lamp over it and put a thermometer in it (do the*
same with the large one). 2.) Record temperature 3.) Place the large and small dog house on ice.
4.) Follow step 2.

In the Doghouse

Annotation of the 2-point response:

Investigation Attributes	**Value Point**	**Annotation**
Prediction	0	No specific prediction is given
Prediction Reason	0	No reason
Materials	0	No timer
Controlled (kept the same) Variable	1	*1 lamp*
Changed (manipulated) Variable	1	Large and small doghouse
Measured (responding) Variable	1	*Record temperature*
Record Measurements	1	*Record temperature*
Trials are Repeated	0	Trials are not repeated
Logical Steps	0	Steps are too vague to repeat. No times are given.
Total	**4**	**2** Score Points

In the Doghouse

Annotated example of a 1-point response:

5 Dalia wondered if a large, well-insulated doghouse would stay warmer in winter than a small one. She asked, “How does the size of a well-insulated doghouse model affect the inside temperature of the doghouse model?” Plan an investigation to answer Dalia’s question.

Be sure to include:

- Hypothesis (prediction) of the investigation results
- Procedure that includes:
  - logical steps to do the investigation
  - one controlled (kept the same) variable
  - one manipulated (changed) variable
  - one responding (dependent) variable
  - how often measurements should be taken and recorded.

Question: How does the size of a well-insulated doghouse model affect the inside temperature of the doghouse model?
Hypothesis (prediction): *I think the bigger the better for a* <u>*well-insulated*</u> *doghouse model*
because It traps more heat inside.
Materials: *Doghouse models (big) (small) w/insulation , thermometer, timer*
Procedure: *1) Place a thermometer in the big doghouse model. 2) Put the model in the freezer.*
3.) Set time for 4 hours. 4.) Repeat steps 1-3 for small doghouse model.

In the Doghouse

Annotation of the 1-point response:

Investigation Attributes	**Value Point**	**Annotation**
Prediction	0	*...the bigger the better* cannot be interpreted to mean the inside temperature.
Prediction Reason	1	*...because it* (bigger) *traps more heat inside*
Materials	0	No freezer
Controlled (kept the same) Variable	1	Same freezer
Changed (manipulated) Variable	1	Big and small doghouse models
Measured (responding) Variable	0	Nothing is measured
Record Measurements	0	None
Trials are Repeated	0	Trials are not repeated
Logical Steps	0	Cannot be effectively repeated because nothing is measured.
Total	**3**	**1** Score Points

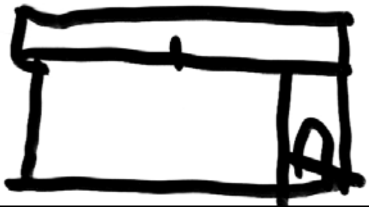
In the Doghouse

Annotated example of a 0-point response:

5 Dalia wondered if a large, well-insulated doghouse would stay warmer in winter than a small one. She asked, "How does the size of a well-insulated doghouse model affect the inside temperature of the doghouse model?" Plan an investigation to answer Dalia's question.

Be sure to include:

- Hypothesis (prediction) of the investigation results
- Procedure that includes:
  - logical steps to do the investigation
  - one controlled (kept the same) variable
  - one manipulated (changed) variable
  - one responding (dependent) variable
  - how often measurements should be taken and recorded.

Question: How does the size of a well-insulated doghouse model affect the inside temperature of the doghouse model?
Hypothesis (prediction): *The bigger it is the colder it would get*
Materials: *large shoe box, foam, thermometer, ice*

Procedure: *Put ice around the box then put it in the freezer, the ice will be like snow + the freezer*
Will be like the cold air

In the Doghouse

Annotation of the 0-point response:

Investigation Attributes	**Value Point**	**Annotation**
Prediction	1	*The bigger it* (doghouse model) *is the colder it* (doghouse model) *would get*
Prediction Reason	0	No reason
Materials	0	No freezer
Inappropriate procedure – Procedure is extremely vague (Note 1d)		
Controlled (kept the same) Variable	0	
Changed (manipulated) Variable	0	
Measured (responding) Variable	0	
Record Measurements	0	
Trials are Repeated	0	
Logical Steps	0	
Total	**1**	**0** Score Points

APPENDIX C

THE PROGRESSION FROM NOVICE TO EXPERT
IN THE SCIENCE INQUIRY EALR

Appendix C

The Progression From Novice to Expert in the Science Inquiry EALR

Grade Level Expectation (GLE) IN02 2.1.2 Planning and Conducting Safe Investigations

Grades K-2

Understand how to plan and conduct simple investigations following all safety rules.

Evidence of Learning:

Classroom Only:

- Make observations and record characteristics or properties.
- Make predictions of the results of an investigation.
- Plan and conduct an observational investigation that collects information about characteristics or properties.
- Collect data using simple equipment and tools that extend the senses (e.g. magnifiers rulers, balances, scales, and thermometers).
- Follow all safety rules during investigations.

Grades 3-5

Understand how to plan and conduct simple investigations.

Evidence of Learning:

WASL and Classroom:

Given a description of a scientific investigation, items may ask students to:

a) Make a prediction related to the given investigation question.
b) Identify one of the variables kept the same (controlled) in the given investigation.
c) Identify the one variable changed (manipulated) in the given investigation.
d) Identify the measured (responding) variable in the given investigation.
e) Make a logical plan for a similar new investigation for a new investigation question with a different changed (manipulated) and/or measured (responding) variable. A logical plan includes step by step instructions clear enough that others could do the investigation.
f) Identify or describe simple materials, equipment, and tools (e.g. magnifiers, rulers, balances, scales, and thermometers) for gathering data and extending the senses.
g) Identify or describe ways to record and organize observations/data from multiple trials or long periods of time using data tables, charts, and/or graphs.
h) Identify or describe safety rules for an investigation.

Classroom:

- Plan and conduct a simple investigation.
- Use simple materials, equipment, and tools (e.g. magnifiers, rulers, balances, scales, and thermometers) for gathering data and extending the senses.
- Gather, record, and organize observations/data from multiple trials using data tables, charts, and/or graphs.
- Identify and use units of measure appropriate for the investigation.
- Follow safety rules during investigations.

<u>Grade Level Expectation (GLE) IN02 2.1.2 Planning and Conducting Safe Investigations</u>

<u>Grades 6-8</u>

Understand how to plan and conduct scientific investigations.

Evidence of Learning:
WASL and Classroom:
Given a description of a scientific investigation, items may ask students to:

a) Make a prediction (hypothesis) related to the investigation question and give reasons for the prediction.
b) Identify one of the controlled variables (kept the same) in the given investigation.
c) Identify the manipulated (changed) variable in the given investigation.
d) Identify the responding (dependent) variable in the given investigation.
e) Make a logical plan for a similar new investigation for a new investigation question with a different manipulated (changed) and/or responding (dependent) variable. A logical plan includes step by step instructions clear enough that others could do the investigation.
f) Identify or describe appropriate materials, tools, and/or computer technology to gather data for an investigation.
g) Identify or describe ways to record and organize observations/data from multiple trials and/or long periods of time using data tables, charts, and/or graphs.
h) Identify, describe, and/or explain safety requirements for an investigation.

Classroom:

- Plan and conduct a scientific investigation.
- Use appropriate materials, tools, and available computer technology to gather data for the investigation.
- Gather, record, and organize observational data from multiple trials using data tables, charts, and/or graphs.
- Identify and use units of measure appropriate for the investigation.
- Follow safety requirements during investigations.

Grade Level Expectation (GLE) IN02 2.1.2 Planning and Conducting Safe Investigations

Grades 9-10

Understand how to plan and conduct systematic and complex scientific investigations.

Evidence of Learning:

WASL and Classroom:

Given a description of a scientific investigation, items may ask students to:

a) Make a hypothesis (prediction with cause-effect reason) related to the investigative question.

b) Identify or describe two of the controlled variables (kept the same) in the given investigation.

c) Identify or describe the manipulated (changed) variable in the given investigation.

d) Identify or describe the responding (dependent) variable in the given investigation.

e) Make a logical plan for a similar new investigation with a new investigative question with a different manipulated and/or responding variable. A logical plan includes step by step instructions clear enough that others could do the investigation.

f) Identify or describe appropriate materials, tools, and techniques, including mathematical analysis and available computer technology, to gather and analyze data.

g) Identify or describe ways to record observations/data from multiple trials and/or long periods of time using data tables, charts, and/or graphs.

h) Identify, describe, and/or explain safety requirements for an investigation

i) Identify or describe an experimental control condition. (An experimental control condition is an unchanged condition that is used to insure the manipulated variable caused the changes in the responding variable when investigating complex systems.)

j) Identify or describe validity measures, in addition to controlled and manipulated variables, for an investigation. (Validity means that the investigation answered the investigative question with confidence; the manipulated variable caused the change in the responding or dependent variable.)

Classroom:

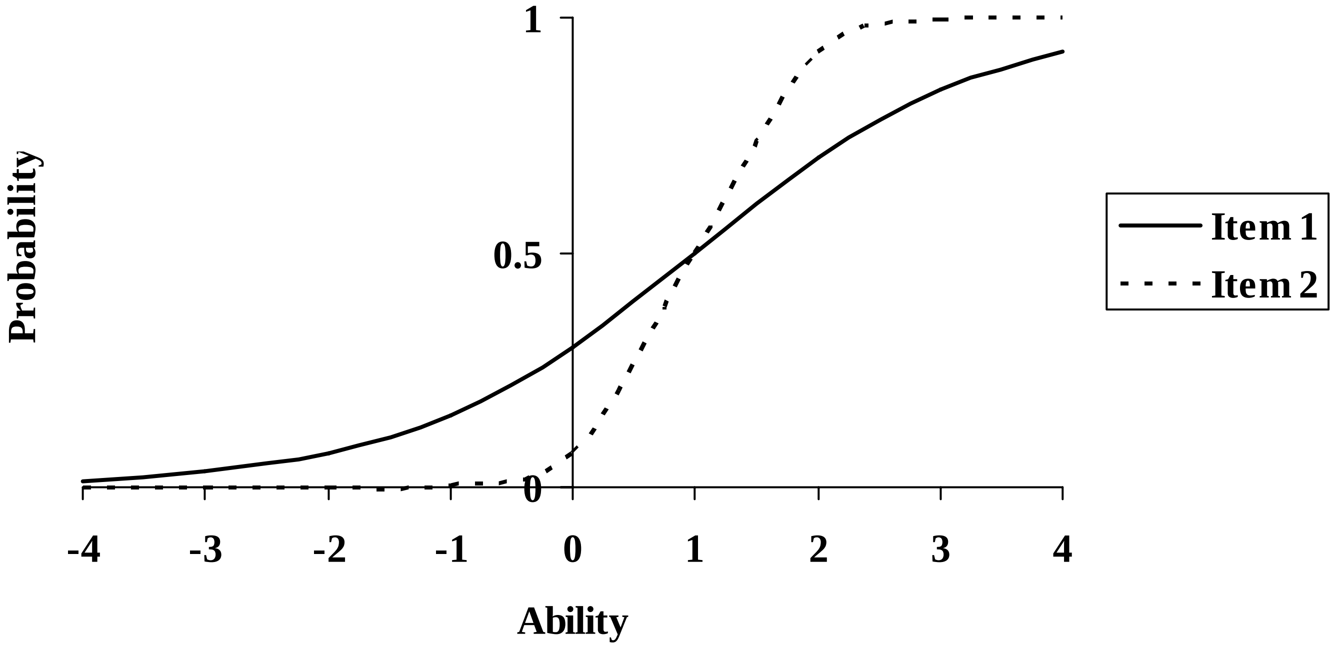
- Plan and conduct a complex scientific investigation.
- Use appropriate materials, tools, and techniques including mathematical analysis and available computer technology, to gather and analyze data.
- Gather, record, and organize observational data from multiple trials using data tables, charts, and/or graphs.
- Identify and use appropriate units for the investigation.
- Follow safety requirements needed in an investigation.
- Set up an experimental control condition when appropriate in an investigation.
- Use validity measures, in addition to controlled and manipulated variables, in an investigation.

APPENDIX D

ITEM CHARACTERISTICS CURVE FOR TWO ITEMS

Appendix D

Item Characteristics Curve for Two Items

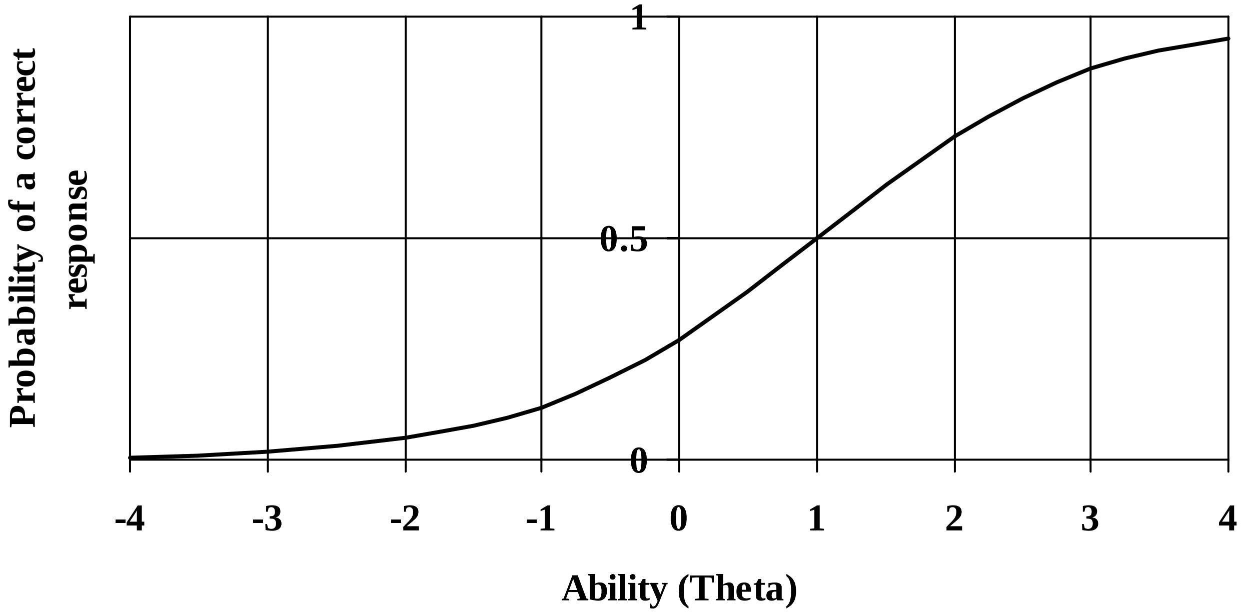


APPENDIX E

POSITIVE ITEM DISCRIMINATION

Appendix E

Positive Item Discrimination

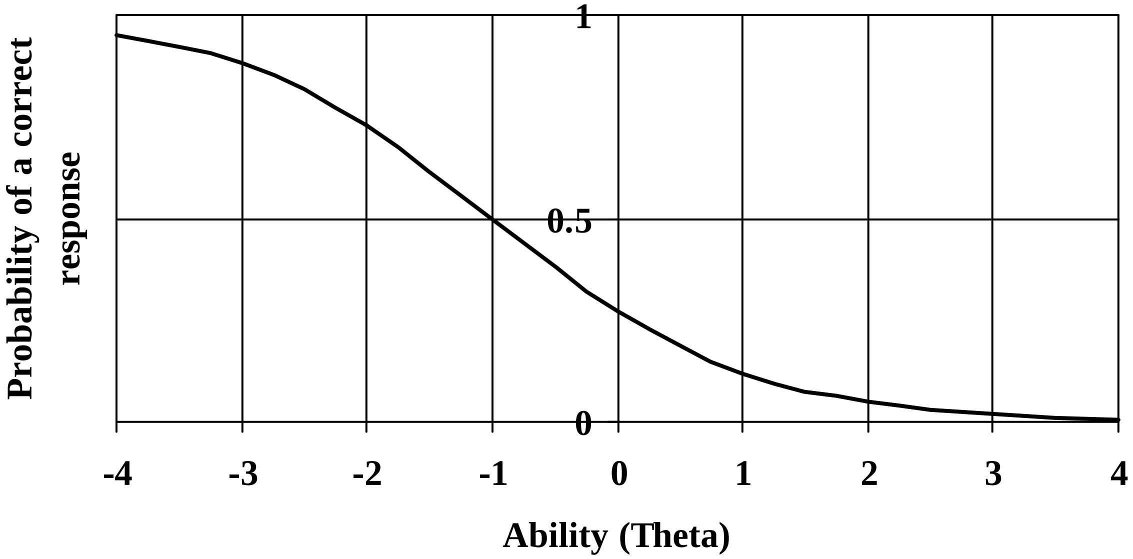


APPENDIX F

NEGATIVE ITEM DISCRIMINATION

Appendix F

Negative Item Discrimination



APPENDIX G

CONSTRUCTS MEASURED ON “IN THE DOGHOUSE” SCENARIO

Appendix G

Constructs Measured on "In the Doghouse" Scenario

In the Doghouse							Grade: 8			
Description: Students investigate the effect of insulating a doghouse model with different materials on the inside temperature of the doghouse model.										
Item Description		**EALR Strand and Grade Level Expectation for Construct**					**Item Type**			
		Properties of Systems	**Structure of Systems**	**Changes in Systems**	**Inquiry in Science**	**Design of Solutions**	**M/C**	**Short Response**	**Extend Response**	**Cognitive Level**
1	Apply an understanding of how to conduct scientific investigations by identifying the responding variable.				***IN02 2.1.2 d***		B			I
2	Analyze solutions to a problem by identifying how to change a real doghouse to create warmer temperatures inside due to sunlight.					***DE03 3.1.3 d***	C			II
3	Identify how the parts of a solar-powered fan interconnect and influence each other.		***ST01 1.2.1 a***				B			I
4	Analyze solutions to a problem by explaining how to change the entrance to the doghouse to keep more heat energy inside.					***DE03 3.1.3 d***		SA		II
5	Apply an understanding of how to plan a scientific investigation.				***IN02 2.1.2 e***				ER	II
					Total		**3**	**1**	**1**	I=3pts. II=6pts.
					Ideal Totals		3-6	1-2	0-1	I: 25% II:75%