The Dissertation Committee for Erica Rae Slate Certifies that this is the approved version of the following dissertation:

An Examination of the Validity of the Mathematics Exit Level Texas Assessment of Knowledge and Skills

Committee:

Jill A. Marshall, Supervisor

Guadalupe D. Carmona

Edward J. Fuller

Susan H. Hull

Jennifer C. Smith

An Examination of the Validity of the Mathematics Exit Level Texas Assessment of Knowledge and Skills

by

Erica Rae Slate, B.S.; M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2005

Dedication

To my husband, Wes Young, for his unending patience, love and support.

To my mother, Lynda Stanbery, for her guidance, support and unconditional love.

And to the loving memory of my father, Raymond Slate.

Without the love and encouragement I received from you all, this work would not have been possible. Thank you.

Acknowledgements

There are many people whom I would like to thank for many reasons. First and foremost, I would like to thank my adviser, Dr. Jill Marshall. I am very fortunate to have had the opportunity to work with her for the last few years. She has been an excellent role model throughout my time at the University of Texas at Austin; constantly leading by example. Watching her for the last four years, I have developed a sense of the kind of educator and researcher I hope to become. She spent countless hours reading my drafts and discussing ideas with me. When I would get stuck and not know how I wanted to proceed or I could not figure out how to say what I wanted to say, without fail, she would come to my rescue with some insight into my dilemma and all would be right again. She has been not only a wonderful advisor, but a friend as well. She was always willing to help in any way she could, which was invaluable after I moved out of the state. She would help me return library books, turn in forms and would even loan me her car when I would come back into town. I look forward to working with her as a colleague in years to come. Without her untiring support and encouragement, I feel sure I would not be where I am today. There is no way I can truly express the gratitude I feel toward her for everything she has done. I can only hope that one day I will have a chance to repay her kindness.

I am also grateful to the other members of my committee, Dr. Jenn Smith, Dr. Lupita Carmona, Dr. Susan Hull and Dr. Ed Fuller, for their time and effort devoted to my project. Their input proved to be invaluable first as I developed my study and then as I analyzed my results. Each member contributed his or her own expertise. I am so very fortunate to have been able to work with each of them.

I would also like to thank Dr, Jere Confrey for her leadership early in my graduate career at UT. Working with her as a graduate research assistant gave me an opportunity to experience some extraordinary things, not the least of which was helping her with her work as the chairperson for the Committee for a Review of the Evaluation Data on the Effectiveness of NSF-Supported and Commercially Generated Mathematics Curriculum Materials. Through her work I was exposed to many different projects and issues that

otherwise I would never have seen. I know my graduate school experience was greatly enriched by the work I did with her and I appreciate the time and energy she devoted to my development as an educator and a researcher.

I wish to also thank several people from earlier in my educational career. I am indebted to Dr. Bill Bauldry for his encouragement and support during my undergraduate and early in my graduate career. He has always believed in my abilities and has encouraged me to pursue a higher degree. The advice he has given me over the years has always led me to the right path. Dr. Deborah Crocker, my Master's thesis adviser, was one of the first people to get me interested in Mathematics Education research in the first place. I appreciate the time and energy she invested in me. Finally, I would like to thank the late Dr. Jimmy Smith. He is the reason I pursued a Master's degree in Mathematics Education. At my undergraduate graduation, he came up to me and offered me a graduate teaching assistantship and basically would not take no for an answer. I was set to begin a high school mathematics teaching career but decided to take him up on his offer instead.

I would be remiss not to mention the content area specialists who participated in my research. I truly appreciate the time you all devoted to my project. I know you all have very busy lives and I am very grateful for the work you did. Also, the teachers at Connally High school for letting me disrupt their classes to recruit students for my interviews. I would like to especially thank Bill Humphries for helping me get my foot in the door, so to speak, with the teachers. He also let me use part of his classroom for a home base while I was conducting the interviews. Certainly my work would not have been possible without his help.

I cannot mention the teachers at Connally High school without also thanking the students. I could not have asked for a better group of volunteers. Each one of the students provided me a wealth of information. I thank them for taking time out of their day to meet with me and for taking the interviews seriously, which enabled me to get so much good information.

I would also like to thank my graduate school friends: David, Katie, Sibel and Lewis. I know that I would not have made it through without your friendships. David was

my true advisor, helping me select the right classes every semester and in general showing me the ropes. Katie has always been a good friend, ready to listen to any ideas or problems I may have had, whether they were academic in nature or not. I am lucky to have had good friends like Sibel and Lewis as well. We spent many hours working on various projects, Sibel and I especially, and I appreciate you both. I am fortunate to have gotten to work closely with all of you. I respect you all as colleagues and as friends.

On a more personal note, I would like to thank my best friend Janet Novoselich for always being there. She helped me by just being a good listener and encouraging me to push on through when things would get tough. I would also like to thank her husband, Brian, who was always willing to run around campus and help me out when I needed it. I could not ask for better friends.

I want to also lovingly thank my parents, Lynda Stanbery and the late Raymond Slate. Both of my parents have always encouraged me to continue my education and without their love and support I would certainly not be where I am now. My mother, especially, instilled in me the importance of higher education and without her encouragement I would not have been able to make it through all of the long years of school. She has always encouraged me in my educational pursuits, even when they took me far from home. She is a great role model and I truly appreciate her unconditional love and support.

Most importantly, I want to acknowledge my husband, Wes Young, for his help throughout my doctoral program. I cannot imagine having a more supportive and loving husband. He has put in countless hours of proofreading and editing, as well as taking over household duties while I have been focused on my work. He has put up with me taking over the dining room table with my papers and files for weeks on end. He also has kept me on track and never let me even consider giving up. Although I am so very fortunate to have had him in my life as a friend for the past 25 years, I am especially lucky to have had him for a husband for the last three. I am forever grateful to you for everything!

An Examination of the Validity of the Mathematics Exit Level Texas Assessment of Knowledge and Skills

Publication No._____

Erica Rae Slate, PhD.

The University of Texas at Austin, 2005

Supervisor: Jill A. Marshall

This study examines the validity of the Spring 2004 Mathematics Exit Level TAKS. In particular, I examined the test through three forms of evidence: content area specialist surveys, statistical analysis of item-level data from 4340 students provided by TEA, and individual interviews conducted with thirty-four 11th grade students. These multiple lines of evidence give a clear understanding of the actual constructs the test measures.

Because of the high-stakes nature of this exam, it is important to examine its validity closely. Assessing the validity of a test involves looking at the appropriateness, the meaningfulness and the usefulness of the test through an empirical investigation into the underlying constructs. Each aspect of this study has provided insight into the underlying constructs of the TAKS. TEA states ten broad objectives TAKS is supposed to cover. Each objective is further broken down into detailed sub-objectives. The TEA-stated objectives and sub-objectives are used in this study as the intended constructs of

the test. In general, the content area specialists' surveys did not confirm the TEA-stated objectives for the test, however this could be due to nuances in the way TEA defined various sub-objectives.

A factor analysis was conducted on the TEA data set in order to see if the items would factor along the TEA-stated objectives, however since the test is designed to be multidimensional, it is not surprising that the items did not factor along objectives. Differential Item Functioning (DIF) was also conducted in order to determine if any items were particularly problematic for various subgroups. Significant DIF was detected in almost one-fourth of the items usually with African American students as the disadvantaged group.

The most useful information came from the student interview data conducted on twenty of the items. Through these interviews, the true constructs the items measure were revealed. In some cases the student interviews validated the TEA-stated objectives, however in many cases, the student interviews showed a different construct. It is mostly due to the results of the student interviews that I was not able say the exit level TAKS is a valid measure of the intended constructs.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The purpose of this study is to examine the validity of the mathematics portion of the Spring 2004 Exit Level Texas Essential Knowledge and Skills (TAKS) exam. In particular, I examined the test through three forms of evidence: content area specialist surveys, statistical analysis of item level data and individual interviews with students. Through these multiple lines of evidence, I hope to gain a clear understanding of the constructs that the test measures. Messick (1989) clearly defines validity as a unified rather than a partitioned concept. He also establishes what validation studies should entail. It is Messick's definition and framework that I have used in conducting my research.

TAKS Background Information

In 2003 the Texas Education Agency (TEA) adopted the Texas Assessment of Knowledge and Skills as the state-mandated accountability measure, replacing the Texas Assessment of Academic Skills (TAAS). The TAKS is administered each spring to Texas public school students from 3rd through 11th grades. The 11th grade administration is considered the exit level exam which students must pass in order to receive a high school diploma; this exam will be referred to henceforth as the Exit Level TAKS. The Exit level TAKS has four sections: English, Mathematics, Social Studies and Science and is administered over four days in mid-April with each section administered on a different day. The test is not timed therefore each student may take as long as he or she needs.

The mathematics portion of the Exit Level TAKS consists of sixty questions covering content up to and including Algebra I and Geometry. Students are required to have a graphing calculator available to use for the duration of the test. Students are also provided with a formula sheet and measurement conversion chart (see appendix A). The test covers a set of TEA-stated objectives, which remain constant from year to year based on the Texas Essential Knowledge and Skills (TEKS) curriculum. The objectives are designed to "serve as heading(s) under which the TEKS can be meaningfully grouped"

(TEA, 2004a, p. 2). The questions are predominately multiple-choice with a limited number of free-response, griddable items. The free response items are designed to "allow students to work on a problem and determine the correct answer without being influenced by answer choices" (TEA, 2004b, p. 3).

The TAKS program was developed over three years and included "input from Texas teachers, administrators, parents, members of the business community, professional education organizations, faculty and staff at Texas colleges and universities, and national content-area experts" (TEA, 2004b, p. 1). Each year new items are field-tested. Prior to field-testing, items are reviewed and revised by educators. After field-testing, items are reviewed by the educators once again, along with the corresponding student response data. For the high school mathematics tests, a content validation review is also conducted, due to the "advanced level of content being tested" (TEA, 2004b, p. 2). The final phase of the test development process was the publishing of the Technical Digest. This digest contains "all relevant psychometric, statistical, and historical information needed to judge the quality of the assessments used" (Smisko, et al., 2000, p. 341) and also a detailed description of the test development process. The Technical Digest also contains statistical data such as estimated reliability and validity as well as the standard error of measurement for each objective, and also for subgroup classifications (Smisko et al., 2000).

Each year after the administration of the TAKS, TEA publishes school and district "report cards" through the Academic Excellence Indicator System (AEIS). The AEIS reports are available to the public through TEA and include detailed information on each school such as TAKS pass/fail rates for all students as well as various disaggregated subgroups, attendance and dropout rates, teacher and staff background data, and budgeting and expenditure data.

Rationale for this Study

Because of the high-stakes nature of the Exit Level TAKS exam, it is important to examine its validity closely. According to *High Stakes: Testing for Tracking, Promotion*

and Graduation, measurement validity, the degree to which a test measures the content domain tested, should be examined in order to determine whether a test is being used appropriately (Heubert & Hauser, 1999). Barbara Plake states that high-stakes tests that are used to make important decisions can also be useful in informing instruction, but only if the "technical quality of these tests [is] adequate to support these purposes" (Plake, 2002, p. 145). In their work, Albrecht and Joles (2003), Sloan and Kelly (2003), and Valencia et al. (2001) all call for a close examination of the validity of high-stakes tests.

Research Questions

This study is designed to answer one overarching research question: Overarching question: Is the 2004 TAKS Exit Level Math a valid measure? In order to answer that question, I investigated three specific questions:

1) How do content area specialists characterize the items in relation to the content covered and the skills needed to solve the items?
2) Is there statistical evidence that certain items are problematic for students or groups of students?
3) Are there underlying factors that explain student responses to items?

Although there is a great body of work that has been conducted on high-stakes tests, there is very little that examines the multiple lines of evidence included in my design. By conducting surveys of the content area specialists, I hope to reveal the underlying objectives measured by the test. In the statistical analysis of the item-level data I hope to identify specific questions that seem to be particularly easy or difficult for the students as a whole, as well as for various subgroups of students. The individual student interviews on the test should help me to identify some of the underlying processes the students use to solve the problems. Each of these methods conducted separately would yield interesting information about the test; however, by examining all three together, I should be able to develop a clearer understanding of the validity of this test.

Chapter 2: Review of Literature

Test Validity

When test validity standards originally were codified in the 1954 Technical recommendations for psychological tests and diagnostic techniques, published by the American Psychological Association (APA), validity was partitioned into four types: content, predictive, concurrent and construct (Shepard, 1993). Content validity was applied to tests that measured a person's performance "on a defined universe of tasks" (Shepard, 1993, p. 408). Both predictive and concurrent validity involved an external criteria; measuring predictive validity required collecting data after the test was conducted, and concurrent validity was used to determine if two tests measured the same material. Construct validity was used for tests that measured "unseen traits such as intelligence or anxiety" (Shepard, 1993, p. 409). In 1966, the APA revised this publication and collapsed two of the above types, predictive and concurrent, into one type: criterion-related validity, thus creating the "holy trinity," a term used by Guion as referenced by Shepard (1993). In the decades since, this definition of validity has evolved through the work of individuals such as Guion, Landy, Tenopyr and Cronbach, and most notably Messick, who is credited with redefining the concept of validity.

In his definitive chapter on validity in *Educational Measurement* (1989), Messick described a unified notion of validity where instead of three types of validity, one looks at validity holistically. He begins by defining validity as "an integrated evaluative judgment to which empirical evidence and theoretical rationales support the *adequacy* and the *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (Messick, 1989, p. 13). Validating an inference involves analyzing the degree to which various forms of evidence support the inference, as well as showing that alternative forms of evidence do not. Validating an action based on a test score involves validating the meaning of the score as well as the value of the action's consequences. Throughout his chapter, Messick continues to establish the theoretical basis for this framework as well as discussing the evolution of the concept of validity in relation to the

progression of the philosophy of science that was associated with various definitions of validity. In this review, I will outline some of Messick's key points as a means to further classify what conducting a validity study should entail.

One of the most important aspects of Messick's conception of validity is that it is not a discrete variable to be measured and reported. Validity is measured in degrees, not absolutes. Also, validity evidence is ever-changing and evolving. As he states, "existing validity evidence becomes enhanced (or contravened) by new findings…validity is an evolving property and validation is a continuing process," (Messick, 1989, p. 13). It is also important to note that tests themselves do not have validity outside the context of measurement. Essentially, validity is a measure of test responses rather than the tests themselves; "what is to be validated is not the test or observation device as such but the inferences derived from test scores or other indicators," (Messick, 1989, p. 13).

Validation of a measure must also take into account multiple sources of evidence, depending on what type of inferences are to be made from the results of the measure. Evidence, as Messick defines it, is not just a collection of facts. He quotes a passage from *The methodology of the social sciences* (1949) by Weber, saying, "empirical data are always related to those evaluative ideas which alone make them worth knowing and the significance of the empirical data is derived from these evaluative ideas," (Weber, as quoted in Messick, 1989, p. 16). Fact, meaning and values are intertwined, essential elements in the validation of a measure. The key elements in a validation study are "the interpretability, relevance, and utility of scores, the import or value implications of scores as a basis for action, and the functional worth of scores in terms of social consequences of their use" (Messick, 1989, p. 13)

Rather than the traditional methods of partitioning validity into three distinct types, Messick seeks to describe validity as a unified, yet faceted, concept. Validity has two main facets: the source of justification of the testing, either through evidence or consequence, and the outcome of the testing, either interpretation or use (Messick, 1989). Both of these facets are interrelated, therefore, Messick represents them as a 2x2 matrix, which is reproduced below:

	Test Interpretation	Test use
Evidential basis	Construct Validity	Construct Validity + Relevance/utility
Consequential basis	Value implications	Social consequences

Figure 2.1: Facets of validity, reproduced from Messick (1989)

Messick (1989) offered the representation as:

> A way of cutting and combining validity evidence that forestalls undue reliance on selected forms of evidence, that highlights the important though subsidiary role of specific content-and criterion-related evidence in support of construct validity in testing applications, and that formally brings consideration of value implications and social consequences into the validity framework. (p. 20)

The traditional three-part division of validity evidence fosters the notion that only one type of evidence is needed to evaluate the validity of any particular test, which would be insufficient in Messick's model. The above matrix is intended to show the interconnectedness between the facets rather than imply distinct boundaries. The distinctions are "fuzzy because they are not only interlinked but overlapping" (Messick, 1989, p. 20). The reason for this fuzziness, so to speak, is that validity is a unitary concept, not meant to be partitioned. Messick states that this matrix representation does nonetheless help the reader to visualize some of the complexities and nuances in the validation of test scores (Messick, 1989). Each of the cells in the matrix represents an aspect of validity, which should be examined both in its own right and also in relation to the other cells.

The evidential basis of test interpretation is made up of the evidence related to score interpretation based on the construction of the test. It is what some would call the technical quality of the test. Here, Messick states, one must be careful of two main threats to validity: construct under-representation and irrelevant test variance. Construct under-representation refers to a situation where a test is too narrow and does not include key aspects of the construct measured. Irrelevant test variance will occur if test questions contain extraneous hints or clues not related to the construct, such as grammatical inconsistencies between the stem of the question and the incorrect answer choices. Irrelevant test variance can also occur from test taking strategies such as one observed by McNeil (2000), "Three in a row? No, no, no!" referring to the unlikelihood of a multiple-choice test having the same answer choice three times in a row (p. 730).

The consequential basis of test interpretation includes the relationship between the value implications associated with constructs and the measures associated with them. These value judgments affect not only the score interpretation but also the consequences attached to the test scores (Messick, 1989). For example, if value were placed on the ability to solve contextual mathematics problems, we would include such problems on an assessment. If we decide that a high school graduate should be able to solve these contextual problems and a high school student fails to demonstrate proficiency, we would not let that high school student graduate. Values affect how tests are constructed in that "values form a basis for the identification and selection of problems and for the priorities and resources allocated to their solution" (Messick, 1989, p. 59). These values carry over into score interpretation as well, which make them an important part of test validity. As the author states, "value implications of score interpretation are not just a part of score meaning, but a socially relevant part that often triggers score-based actions and serves to link the construct measured to questions of social policy" (Messick, 1989, p. 63).

The evidential basis for test use is similar to the evidential basis of test interpretation in that it is partially defined by construct validity. However, in relation to test use, another factor is added: the relevance of the scores to the purpose of the test. Here is where one would evaluate the justification for the test use by examining the

relevance and utility of the test scores, as well as the score meaning(s), based on a range of evidence. One method of collecting evidence, Messick says, is to have a group of experts judge the content of test items based on curricular or instructional relevance. In other words, do the items relate to the objectives of the curriculum being tested? Curricular relevance is used as a broader term encompassing the curriculum as a whole, while instructional relevance refers to only that which the students were actually exposed to (Messick, 1989). The expert analysis alone, however, is not sufficient, according to Messick. One must look at other forms of evidence as well, "especially evidence to discount the operation of construct-irrelevant method variance" (Messick, 1989, p. 70).

The consequential basis of test use incorporates the social consequences of testing as an integral part of validity. Here, one should look at the appropriateness of the test administered, as well as, the unintended outcomes of the test. The unintended outcomes should be examined especially closely if it is determined that the unintended outcomes can be tied to sources of test invalidity. Even if no such ties exist, Messick states that it is still important to evaluate the consequences and side-effects of test use (Messick, 1989). He uses a rather illustrative analogy of drug testing to show the relevance of this aspect of validity. It is not wise to take a drug without examining the potential side-effects; the pain may go away, but who knows what else may happen. According to Messick, administering a test should be viewed in much the same way. If it is determined that an unintended outcome is related to variance due to some flaw in the test (e.g. bias or construct irrelevant test variance), then the unintended outcome is an issue of invalidity. On the other hand, if it is determined that the unintended outcome is related to a valid property of the construct being measured (i.e. actual difference in ability level), then that is considered an issue of test validity and at that point the unintended consequence becomes an matter of social policy (Messick, 1989).

Each of these four aspects of validity is important to consider in any validity study. These elements can be examined separately; however, information from each one can help to inform conclusions made on the others. A validity study should take each of these elements into account.

High-Stakes Testing

A high stakes test is an assessment where the results are used to make a high stakes decision. This decision can be for a school district, an individual school and/or teacher, and individual students as well. High stakes decisions for individual students "involve tracking (assigning students to school programs or classes based on their achievement levels), whether a student will be promoted to the next grade, and whether a student will receive a high school diploma" (Heubert & Hauser, 1999, p. 1).

Although testing in schools can be traced back hundreds of years (Allen & Yen, 1979; Ravitch, 2002), what is referred to as high stakes testing in America has become prominent in the last few decades. In 1965, Congress passed the Title I of the Elementary and Secondary Education Act of 1965, introducing large-scale testing as an "integral part of federal support for the education of low-achieving children in poor neighborhoods" (Heubert & Hauser, 1999, p. 15). In 1969, the first National Assessment of Educational Progress (NAEP) was administered to measure the quality of American education (Vinovskis, 1998). The 1970's marked the beginning of the minimum competency testing movement where large-scale standardized tests played a role in holding students accountable (Heubert & Hauser, 1999). The publication of *A Nation at Risk* in 1983 propelled the standards-based reform movement in education and also drew public attention to student performance on tests.

As a result, by the mid-1980's many states began mandating various forms of minimum competency testing. Some states implemented test-based requirements for graduation. By the mid to late-nineties, nearly half of the states had serious consequences for schools tied to student test performance (Heubert & Hauser, 1999). In 2001, President Bush signed into law the No Child Left Behind Act (NCLB), which was a revision of the Elementary and Secondary Education Act of 1965. The NCLB, which emphasized "increased accountability for States, school districts, and schools," (U.S. Department of Education, 2002, p. 1) requires that all states implement standardized testing programs "covering all public schools and students" (U.S. Department of Education, 2002, p. 1).

Through the NCLB, high-stakes testing has become a part of life for American public school students.

The issues surrounding the use of high-stakes tests have been the focus of a multitude of scholarly works. In 1999, the National Research Council published *High Stakes: Testing for Tracking, Promotion and Graduation*, the report from the Committee for Appropriate Test Use. The committee was charged with making recommendations on the appropriate use(s) of the tests that are employed to assess student performance. In the report, the committee focused on "tests with high stakes for individual students" (Heubert & Hauser, 1999, p. 2), but they did recognize the potential consequences for students of high stakes accountability measures of teachers, schools, and school districts and encouraged further research to be conducted (Heubert & Hauser, 1999).

The committee stated three criteria based on professional standards previously established, each of which should be examined in order to determine if a test is being used appropriately. These are measurement validity, attribution of cause, and effectiveness of treatment. The committee defines measurement validity as "whether a test is valid for a particular purpose, and whether it accurately measures the test taker's knowledge in the content area being tested" (Heubert & Hauser, 1999, p. 2). Attribution of cause refers to whether a student's performance on the test in question is actually attributable to the instruction the student received or if it is due to external factors, such as language barriers to understanding. Finally, effectiveness of treatment is defined as whether a test leads to educationally beneficial consequences (Heubert & Hauser, 1999).

Using this framework, the committee developed several principles for appropriate test use in making high-stakes decisions. First, the committee stated that in evaluating test validity, one must look at the specific purpose of the test, because "the important thing about a test is not its validity in general, but its validity when used for a specific purpose" (Heubert & Hauser, 1999, p. 3). Secondly, the committee acknowledges that tests are not infallible and that a single test score is not a precise measure of a student's ability. For this reason, the committee goes on to emphasize that "an educational decision that will have a major impact on a test taker should not be made solely or automatically on the

basis of a single test score. Other relevant information about the student's knowledge should also be taken into account" (Heubert & Hauser, 1999, p. 3). Lastly, the committee observes that tests themselves are not likely to improve educational outcomes in the absence of better school and classroom practices. The committee states that research on practices such as tracking has shown negative outcomes for students in lower tracks, and that retention without appropriate instructional support services is also harmful to students (Heubert & Hauser, 1999).

In addition to this National Research Council report, many researchers have conducted studies focusing on various aspects of high-stakes tests including their intended and unintended outcomes and consequences. One concern among those who discuss the unintended outcomes of high-stakes testing is the narrowing of the curriculum, or in other words, teachers spending inordinate amounts of time teaching to the test. Linda Nathan, the headmistress of the Boston Arts Academy, a three year public high school for the arts, expressed concern in a Phi Delta Kappa article that because of the state's requirement for students to pass the Massachusetts Comprehensive Assessment System (MCAS), the teachers at her school would have to take time away from arts instruction to make time for test preparation exercises. She is also concerned that innovative teachers will be discouraged from developing new curricula for their classes because of concern for student test scores (Nathan, 2002). Other researchers express similar concerns about the narrowing of curriculum in schools as a result of high-stakes testing. Glaser and Silver (1994) state that research shows the narrowing of curriculum occurs disproportionately in classes with a larger proportion of minority students. In reference to the TAAS, McNeil (2000) and Valenzuela and McNeil (2000) express concerns that test preparation programs have replaced the regular curriculum in schools labeled as low-performing. Albrecht and Joles (2003), Sloan and Kelly (2003), Amrein and Berliner (2002), Kohn (2000, 2001) all state the issue of teachers teaching to the test and the subsequent narrowing of the curriculum as potential negative consequence of high-stakes testing programs.

Another prevalent view is that high-stakes accountability programs, particularly ones with a high school graduation exam, will lead to an increase in the dropout rates, especially among minority and economically disadvantaged groups (McNeil, 2000; Kohn, 2000; Albrecht & Joles, 2003; Alexander, 2003; Amrein & Berliner, 2002). When one looks at the data on this subject, there is conflicting evidence.

In a study of Florida high school students in a random sample of 14 districts, Griffin and Heidorn (1996) found no significant differences in the dropout rates of economically disadvantaged groups due to the administration of the Minimum Competency Test (MCT). It should be noted, however, that the dropout rate was already high for this group. The one area where Griffin and Heidorn found a significant increase in dropout rates was for those students who failed the exam but otherwise had good overall academic records. They attribute this result to the fact that these students probably could not deal with the stigma attached to having failed the MCT (Griffin & Heidorn, 1996).

Amrein and Berliner (2002) conducted a study of all states with high-stakes testing requirements for high school graduation. At the time eighteen states had exit exam requirements, however two of those had recently implemented the requirements and therefore did not have the longitudinal data required by the researchers. Of the sixteen states with sufficient longitudinal data, eight showed an increase in drop out rates and ten saw a decrease in graduation rates after the high-stakes exams were implemented.

A smaller scale study conducted by Alexander (2003) looked at the dropout rates for one Texas urban school district and found that the dropout rates lowered between 2000 and 2001 for Hispanic, African American and White students, and that the graduation rates for African American students were higher than any other subgroup. However, the author does not explain if these rates are due to the use of a high-stakes test.

In another study of Texas attrition rates, Haney (2000) found that after the implementation of the graduation requirement of the exit level TAAS, dropout levels increased for all groups, but more so for Blacks and Hispanics (Haney, 2000). Fuller and

Johnson (2001) also looked at dropout rates for Texas students, among other things. They concluded that even though, "the dropout rate for Texas is unacceptably high, especially for African-American and Hispanic students," there is no conclusive evidence that the dropout rates are related to the high-stakes accountability system (p. 278).

Other potential negative outcomes of high-stakes testing mentioned in the literature include an increase in anxiety for students, stigma attached to students who do not pass, resulting in lower self-esteem, and an increase in the attrition rate of teachers, (Kohn, 2000; Albrecht & Joles, 2003; Sloan & Kelly, 2003; Amrein & Berliner, 2002). Data on these outcomes is not presented and would therefore be areas for future research.

Several researchers discuss the potential positive outcomes from high-stakes testing programs. Sloan and Kelly (2003) and Goertz and Duffy (2003) both state the usefulness of well-designed tests in giving students and teachers useful information about student knowledge and skills. This information can be used to identify specific areas a student needs to work on. Such tests may also increase student motivation to work harder and take school more seriously. Another benefit of high-stakes testing programs, such as the Texas system, as stated by Skrla and Scheurich (2001), Skrla, Scheurich and Johnson (2001) and Scheurich, Skrla and Johnson (2000), is the raising of standards for all students. In an accountability model such as the one used in Texas, districts are held accountable for the performance of all subgroups, not just aggregate results. The argument here is that this system forces administrators and educators to pay attention to previously neglected groups, such as economically disadvantaged, African-American and Hispanic students. The authors state that through the accountability system, school leaders "moved the academic success levels and school experiences for children of color and children from low-income homes…out of the dank and hideous basement of failure and invisibility where, prior to state accountability, they had remained undisturbed" (Skrla & Scheurich, 2001, p. 256).

In reviewing the literature on high-stakes testing, one other aspect seems clear. There is a distinct spectrum of opinions on the current practices of high-stakes testing in schools. There are clearly those researchers who are opposed to current high-stakes

testing practices, such as Alfie Kohn, who states, "high-stakes testing, which relies on rewards and punishments to increase scores, creates a system that is unfair as well as destructive to learning," (Kohn, 2000, p. 315). Richard Valencia, Angela Valenzuela and Linda McNeil are others who have made a stand against high-stakes testing practices (McNeil, 2000; McNeil & Valenzuela, 2000; Valencia et al., 2001). On the other hand, there are those researchers who stand squarely in favor of the current practices, such as William Mehrens and Susan Sclafani. Mehrens speaks from the perspective of an external evaluator, but his conclusions are strictly "pro-TAAS," at least from a psychometric perspective. In his analysis of the TAAS program in Texas, he found that it met all standards for reliability and validity of test construction and that the system in Texas is working. He says that "without a requirement like the TAAS, students might graduate without having learned what the state has deemed to be a set of minimum requirements" (Mehrens, 2000, p. 389). Others, like Jim Scheurich and Linda Skrla, are optimistic that the current practice will lead to better student outcomes. They acknowledge the problems with many state accountability systems, but their research has shown positive results for children of color and from low-income homes (Skrla & Scheurich, 2001).

In between the two extremes there is a range of other opinions backed by both positive and negative results. For example, I would describe Ed Fuller and Joseph Johnson as guardedly optimistic in their conclusions. They state that the system in Texas is not perfect but, "state accountability systems deserve more rigorous study by all those concerned about the education of children of color and children from low-income homes" (Fuller & Johnson, 2001, p. 281). Similarly, Goertz and Duffy believe in the potential for positive outcomes from these tests if the tests are carefully designed and are proven to be valid assessments for the content they purport to measure. They state, "well designed assessments and accountability systems can focus attention on schools and students who need the most help…but policy-makers must recognize the limits as well as the promise of such policies" (Goertz & Duffy, 2003, p. 10).

Other researchers, for instance Albrecht and Joles (2003), and Glaser and Silver (1994), state the need for assessment measures and seem to accept that high-stakes testing is inevitable. They are, however, concerned with the unintended negative outcomes such as the narrowing of the curriculum. As Glaser and Silver state, "the intended and unintended effects of an assessment on the ways teachers and students interpret results, frame educational objectives, and allocate time, warrant serious examination," (Glaser & Silver, 1994, p. 413). Albrecht and Joles (2003) also state their specific concerns with the potential for disparate impact testing practices might have on particular subgroups of students, saying:

> Testing practices should be thoroughly examined for disparate impact when a group of students performs differently and if the education decisions based on test scores reflect significant disparities based on race, national origin, gender or disability. (p. 87)

Some researchers, like Trueba (2001), and Griffin and Heidorn (1996), come across as neutral, neither advocating nor denouncing the use of high-stakes tests. Trueba tries specifically to provide an unbiased summary of the issues surrounding high-stakes testing in Texas. In his conclusions, he criticizes those on both sides of the debate by saying, "some of the strongest positions against the TAAS are overly protective and can be interpreted as patronizing. On the other hand, some of the extremely supportive positions can be seen as politically opportunistic," (Truba, 2001, p. 340). Griffin and Heidorn, in their study of dropout rates in schools with high-stakes testing programs, did not show any statistically significant differences in dropout rates among ethnic groups but also could not attribute their findings specifically to the high-stakes test (Griffin & Heidorn, 1996). Whether current high-stakes testing practices are beneficial or harmful to students remains to be seen. One thing is clear and that is that testing practices should be examined closely and carefully, especially when they influence high-stakes decisions.

Chapter 3: Methodology

Introduction

The purpose of this study is to examine the validity of the mathematics portion of the Spring 2004 Exit Level TAKS exam. Specifically, I want to look at the individual questions in order to determine if the test is actually measuring what TEA claims, or if there is some underlying pattern to the student responses. Because of the high-stakes nature of this exam, it is important to examine its validity closely. As stated in *High Stakes: Testing for Tracking, Promotion and Graduation*, measurement validity, the degree to which a test measures the content domain tested, should be examined in order to determine whether a test is being used appropriately (Heubert & Hauser, 1999). Barbara Plake states that high-stakes tests that are used to make important decisions can also be useful in informing instruction, but only if the "technical quality of these tests [is] adequate to support these purposes" (Plake, 2002, p. 145). In their work, Albrecht and Joles (2003), Sloan and Kelly (2003), and Valencia et al. (2001) all call for a close examination of the validity of high-stakes tests.

According to Messick (1989) and Shepard (1993) validity in the past has been studied in a fragmented form, usually divided into content, criterion-related and construct validity. Messick (1989) redefined validity as a unified concept integrating test use, values and consequences (Shepard, 1993). It is this unified theory of validity that I will use as a basis for my framework. Validity, as defined by Messick, is, "an integrated evaluative judgment to which empirical evidence and theoretical rationales support the *adequacy* and the *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (Messick, 1989, p. 13). In order to validate a decision made based on a test score, one must "ascertain the degree to which multiple lines of evidence are consonant with the inference, while establishing alternative inferences are less well supported" (Messick, 1989, p. 13).

There are several sources of these multiple lines of validity evidence: the relationship of the test to the content domain, the internal structure of responses, the ways

in which individuals respond to the test, a survey of the relationships between the test scores and other measures, the longitudinal evidence of the test scores across different groups, settings and instructional treatments, and the social consequences of the intended and unintended outcomes of the test (Messick, 1989). In this study, I will focus on collecting evidence from three of these sources: the relationship of the test to the content domain, the internal structure of the test, and the ways in which students respond to various items on the test. Other researchers have studied and will continue to collect data on the remaining three lines of evidence. As part of the statewide accountability model, TEA conducts annual reviews of student scores and compares performance for all students and various subgroups over time (Smisko et al., 2000). Independent work has also been conducted in this area through Webb and others (2001). Work from researchers such as Amrein and Berliner (2002), Fuller and Johnson (2001) and Scheurich et al. (2000) has surveyed student performance on the statewide assessments compared to other tests such as SAT and NAEP, while others have examined the relationships between the test and high school completion (Alexander, 2003; Haney, 2000; Fuller & Johnson, 2001). The "social consequences" aspect has been examined extensively with regard to the Texas test by Valencia et al. (2000), Valenzuela and McNeil (2000), and Scheurich et al. (2000). While work in each of the preceding areas contributes a portion of the validity evidence, no single study has been conducted to examine all six aspects. This is an area for future work.

In my study I collected evidence for the stated aspects in three phases: a survey of content area specialists in the fields of mathematics and mathematics education, an analysis of item-level test data from TEA and interviews with individual students. The content area specialist surveys provide evidence on the relationship of the test to the content domain. The item-level data analysis provides data on the internal structure of the test, and lastly, the student interviews examine the ways in which students respond to the test. I then synthesize the information from each of the three phases in order to create a detailed inspection of the content validity of this exam.

Content Area Specialist Surveys

I distributed surveys to content area specialists in the fields of mathematics and mathematics education. This group of specialists consisted of university mathematicians and mathematics educators as well as experienced high school mathematics teachers. The high school mathematics teachers are included because they have direct contact with the level of students who take the exam as well as the intended content of the exam. The university-level mathematicians and mathematics educators are familiar with the mathematics content as well as the teaching and learning of mathematics.

Each content area specialist received a copy of the exam along with the survey questions. I asked each respondent to go through the test question by question and record the following information for each item:

Correct response
Primary objective
Skill level of item
Additional comments

According to TEA, the math portion of the TAKS is supposed to cover a set of ten primary objectives. I used these objectives as the list of objectives the content area specialists could select from. I also added an eleventh choice of "objective not found in above list" so the specialists could write in an objective for an item if they felt it was not covered by the given list. For the exact protocol for the surveys, see Appendix A.

The survey also included a free response section where the respondents could make additional comments on particular questions as necessary. I asked the content area specialists to make note of questions for which the correct answer is not listed, where multiple answers would be correct, or questions that stand out for some other reason such as being exceptionally easy or difficult.

This survey data were used to pick out specific questions for further analysis. In addition, the objective categorization data was compared with the TEA-stated objectives in order to examine the perceived constructs versus the ones intended by TEA. The items

that do not align with TEA-stated objectives were closely examined, as well as any items identified by the content area specialists as standing out for some reason.

Item Analysis of TAKS Data

The second phase of my work was a statistical analysis of the internal consistency of the exam. TEA makes public the data collected from the TAKS administrations. I requested item-level data from TEA for the exam in question. Specifically, I received a random sample (generated by TEA) of 5000 students, with item-level student response data for the mathematics portion. With these data I conducted an exploratory factor analysis in order to determine if the questions "load" on specific factors, such as content areas. Factor analysis is a statistical technique that assumes the existence of underlying, independent relationships, or factors, in a data set. Various algorithms are used to identify the factors, or eigenvectors, which can serve as a minimum basis to represent the majority of the data. If there are valid factors, then all, or nearly all, of the data can be represented by linear combinations of these factors. If factors can be identified in the data, they may point toward valid explanatory factors for the student outcomes, and thus inform instruction. For example, if all the problems involving volumes or areas, and only these problems, loaded strongly onto a single factor (had high weightings for that eigenvector), that might indicate that student understanding of area and volume is an important factor and that instruction in this area should be examined.

Grimm and Yarnold (1997) define the difference between exploratory versus confirmatory factor analysis to be "[exploratory factor analysis] finds the one underlying factor model that best fits the data; [confirmatory factor analysis], in contrast, allows the researcher to impose a particular factor model on the data and then see how well that model explains responses to the set of measures…Thus EFA primarily represents a tool for theory building, whereas CFA represents a tool for theory testing" (p. 109).

In order to determine the number of factors (eigenvectors) to retain, I used Cattell's "scree test" as referenced in Grimm and Yarnold (1997). The scree test involves plotting each of the eigenvalues versus the factor number, i.e., how many factors had a

significant loading on that eigenvalue. For a data set in which each point was completely unrelated to any other point, the number of eigenvalues would equal the number of data points and each would have a factor number of "1"; the plot would be a flat line at y = 1. Commonly, there will be a small number of eignvalues with much higher factor numbers, and thus this plot will have a sharp drop off in the beginning and will level off for successive factors. According to Grimm and Yarnold, "the eigenvalues (and corresponding eigenvectors) in the steep descent are retained and the eigenvalues in the gradual descent (including the eigenvalues occurring in the transition from steep to gradual descent) are dropped," (1997, p. 103).

While factor analysis can be useful in identifying factors in a multidimensional data set, in this case it may or may not provide any useful information. The TAKS is designed to be a broad, multidimensional assessment covering multiple objectives. The sixty items on the math portion of the TAKS may not provide enough data in order for a factor analysis to adequately identify specific meaningful factors. However, if items do seem to load on specific factors, that information will be interesting to investigate, especially if those factors do not correlate with the stated objectives for the test. For example, they might hypothetically correlate instead with the level of vocabulary in the problem, or whether the problem involved people in some way.

Along with a factor analysis, I conducted an item analysis for each item. Item analysis can be used to empirically assess the quality of test items but also to identify problematic items using various statistical measures such as p-values and Differential Item Functioning (Varma, 2004). The p-value of an item is the proportion of students who answered the item correctly and can range from 0 to +1. For a multiple-choice test with four choices, one would expect by chance alone 25% of students would get a question correct, in other words the p-value of the question is expected to be p=0.25 if nothing is measured except random guessing. A general rule of thumb to use in order to be confident that a question measures more than random guessing is as follows: the p-value should be halfway between the maximum expected p-value and the p-value for guessing (MEC, 1997). In this case the maximum expected p-value is 1.0, since it is

possible for 100% of students to answer a question correctly, and the p-value for guessing is 0.25, so the optimal p-value for a four-answer multiple-choice question is approximately p=0.63. The p-value is one way to measure the difficulty level of an item since it indicates what proportion of students successfully answered the question correctly; the higher the p-value, the easier the question. The interpretation of the p-value statistic is largely subjective, however the ideal range is said to be between 0.5 and 0.9 (MEC, 1997). For the purposes of my work, I am mostly interested in identifying questions that are the most difficult for students, however it may be interesting to look at the questions that seem to be the easiest as well. Initially I focused on items with p-values less than 0.5, but later included items with a p-value greater than 0.8 for examination as well.

I used Differential Item Functioning (DIF) to identify questions with anomalous results. DIF identifies items where test takers of the same ability have different item scores (Zenisky, Hambleton & Robin, 2003). Usually, DIF is detected using the Mantel-Haenzsel statistic. In order to use the Mantel-Haenzsel statistic, the data in question should be divided into two groups: a reference group to use as a baseline for scores and a focus group whose scores will be compared to the reference group (Angoff, 1993). In the item level data I received from TEA, I have data on student ethnic subgroup, limited English proficient (LEP) status, gender, economic disadvantaged status, Math total score, and TAKS total score. DIF analysis could be conducted on any of these variables; I however chose to focus on potential differences across ethnicity levels. Analysis of the other variables could be done in future work.

The Mantel-Haenzsel statistic is essentially a modified Chi-squared test. I used SPSS to create a 2x2 contingency table for each item of interest to compare proportions of students across score groups in the reference versus the focus group who answered the item correctly and who answered incorrectly. If no DIF exists, the proportions in the reference group versus the proportions in the focus groups should be the same. If a significant difference exists, DIF exists for that item. This is important to examine in reference to validity because, according to Angoff (1993), “a large DIF value would

suggest that the item is measuring an additional construct in one of the groups that may not be relevant to the intended constructs of the test," (p. 18). Linn (1993) states that DIF analyses alone will not prove bias and should be used "in conjunction with judgmental review of test content…by investigations of relationships to external criterion measures for focal and reference groups, and investigations of the construct validity of the measures for these different groups," (p. 351).

Each of the statistical methods described above will help to empirically identify problem items. I used this data along with the survey data I received from the content area specialists to zero in on specific items I wished to interview the students on.

Student Interview Data

The third part of my project involved interviewing individual students on selected questions from the test. I selected students on a volunteer basis from a school district that had a prior performance record that closely matched state averages (See figure 3.1).

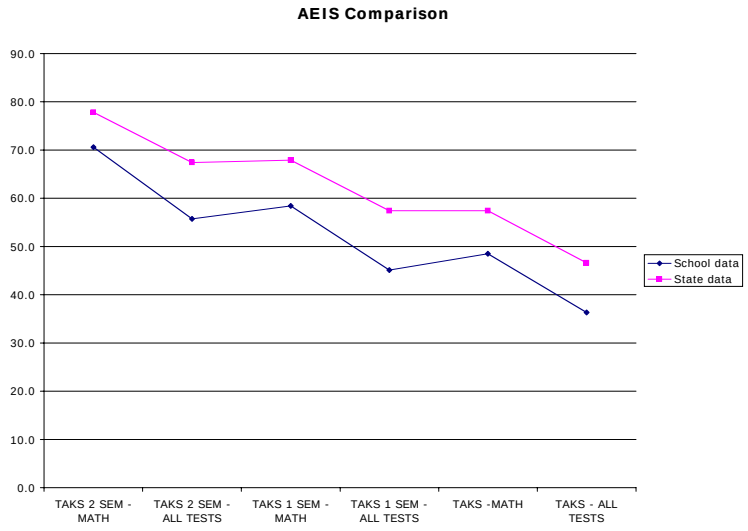


Figure 3.1: Comparison of most recent (2002) Academic Excellence Indicator System (AEIS) available for selected school with state data

In Figure 3.2, below, I show how the scores correlate to one another. I plotted the school data for each measure as the x-coordinate and the state data as the y-coordinate. If the scores were exactly the same, we would expect the resulting scatter plot to fall on the line y=x. The actual resulting line of best-fit shows that the selected school closely matches the state averages for the sub-scores on the 2002 Academic Excellence Indicator System (AEIS). The purpose of this portion of my study is to look at how typical students respond to the questions. Future work could be done using the framework in this study to instead focus on high or low performing subgroups.

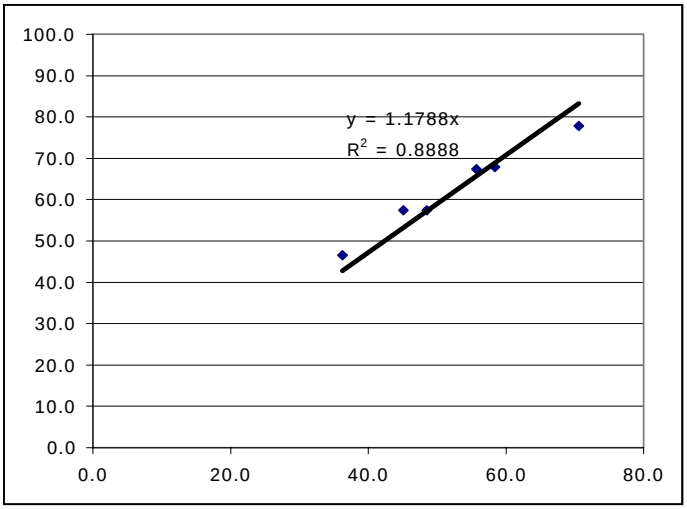


Figure 3.2: Correlation between selected school and state data

Before the interviews were conducted, I obtained Institutional Review Board (IRB) approval through the University of Texas Office of Research Support and Compliance, and I distributed informed consent forms to the students and their parents/guardians. Consent forms were available in English and Spanish, though no students or parents requested the Spanish version of the consent form. Each student interviewed returned both the parental consent and the student consent form.

I interviewed thirty-four 11th grade students who had just taken the 2005 exit level TAKS for the first time, the week prior to the interviews. These students had not taken the 2004 exit level TAKS before; however, since the 2004 items had been released, the

students interviewed may have seen some of the items prior to the interview. The students were recruited from various math and science classes in one central Texas high school. To recruit students I went to several 11th grade level math and science classes, gave a brief description of what the interviews would entail and passed around a sign-up sheet. In order to help attract volunteers, students were told that if they participated they would be invited to a pizza lunch after interviews were completed. All students who volunteered were interviewed. After the initial sign-up, I reviewed the list of volunteers with the math and science teachers whose classes I recruited from in order to ensure I had a heterogeneous sample. I did not do a formal investigation of the students' academic backgrounds; however the teachers stated my volunteer sample included students with a range of mathematical abilities.

The math portion of the exam consists of sixty questions; therefore it was unrealistic to expect a student to sit for hours on end in an extensive interview covering every item. I selected a subset of items identified from the content area specialist surveys and the item analysis described above. Initially, I selected items with p-values less than 0.05 (the most difficult items). I also included the two easiest items, as well as the items the content area specialists identified as being interesting for various reasons.

I spent about 30 minutes with individual students where I asked them to solve several of the TAKS items on paper, talking through their thought processes as they went. I used a "think aloud" protocol as referenced in Ericsson and Simon (1984) where "the subject is specifically asked to vocalize [his] self generated symbols while he is performing his task" (p. 78). In this type of protocol, respondents are asked to work through problems while explaining their thought processes as they go. Below I have included an excerpt from an interview that shows how I initiated each interview.

Interviewer	Well what I am going to have you do is, I've got a whole bunch of different problems here. You don't have to finish all of them, but we'll work through as many as we can and I want you to talk through what you're doing. Write it out, solve it out just like you would on the TAKS but kind of talk to me and tell me what you're

thinking; what you do when you start the problem, how you approach it. Okay?

Respondent Okay. (pause) Do I have to read it?

Interviewer No, you don't have to read it out loud. You can just tell me how you would solve a problem like this and then kind of work it out.

I let the students work through each problem with minimal prompting. I acknowledged statements they made with an "ok" or "alright" in order to allow the students to give a narrative. I would ask for clarification whenever a student made a statement that I felt needed further elaboration. Below is an excerpt from a transcript where the student required further prompting in order to get him to explain his reasoning.

Interviewer So what about this problem?

Respondent I'm trying to remember what this meant. (points to a word)

Interviewer What supplementary means?

Respondent They go like two angles that added up to be 180.

Interviewer Ok

(long pause)

Interviewer So then how are you going to figure out the right answer?

Respondent Hmmm…oh, I got it.

Interviewer Ok. What is it? Tell me.

Respondent Two angles that added up to be 90.

Interviewer Ok

Respondent And seeing that it already has 90 in here,

Interviewer Uh huh. Then what?

Respondent Then I guess it can't be supplementary angles.

I audio taped the sessions and also kept a copy of the students' written work. The students were able to work through, on average, twelve items during the thirty minutes allotted for each interview. After the interviews were completed, I listened to the tapes and made notes on each respondent: which questions they answered, methods they used to solve each problem, and comments they had made. Based on these notes, I classified each problem solution according to a series of category codes using a grounded coding scheme (Strauss & Corbin, 1990). The codes were taken directly from what the students said or did. These codes included, for example, "process of elimination" and "used graphing calculator." As the think-aloud process generates a fairly clear representation of student thinking, it was not deemed necessary to employ a second coder for reliability. Because a fairly limited set of codes were generated in the open coding process and the tasks were fairly constrained, the open coding was sufficient. I then analyzed the codes resulting from the interview transcripts for trends or similarities in solution methods across specific problems. I will discuss some of the trends/patterns I found from the interview data in my results chapter that follows.

The data collected from the student interviews help to explain the underlying constructs of the items as well as why the some of the items are problematic. As Weiss (1994) stated in his book on interview studies, "we can learn also, through interviewing, about people's interior experiences. We can learn what people perceived and how they interpreted their perceptions," (p. 1).

Summary

The purpose of this study is to examine the validity of the mathematics portion of the TAKS test. A more modern definition of validity is that it is not one-dimensional, but rather should be examined through "multiple lines of evidence," (Messick, 1989, p. 13). These lines of evidence include the relationship of the test to the content domain, the internal structure of responses, and the ways in which individuals respond to the test (Messick, 1989). The three phases of this study each provide insight into one of these

aspects of this high-stakes test. The content area specialists provide insight into what objectives the test seems to be measuring. This is especially important if these objectives differ from the objectives the TEA states the exam measures. The examination of the data collected from the test administration is useful in looking at the big picture, so to speak. Analysis of these data provides insight into which questions seem to be systematically problematic to the student population at large, as well as for specific subgroups. Finally, the student interviews provide important information as well. The students are the stakeholders for this exam so how they perceive the questions is relevant to the test's validity. In a multiple-choice exam, the test-takers' thought processes are not taken into account. The interviews with students help to bring some of this into light. Information from each of the three phases will be synthesized in order to create a detailed look at the content validity of this exam.

Chapter 4: Results

Student Interview Data and Content Area Specialist Surveys

Thirty-four 11th grade students were interviewed on 20 of the 60 items. Figure 4.1 below shows the breakdown of students by gender and ethnicity.

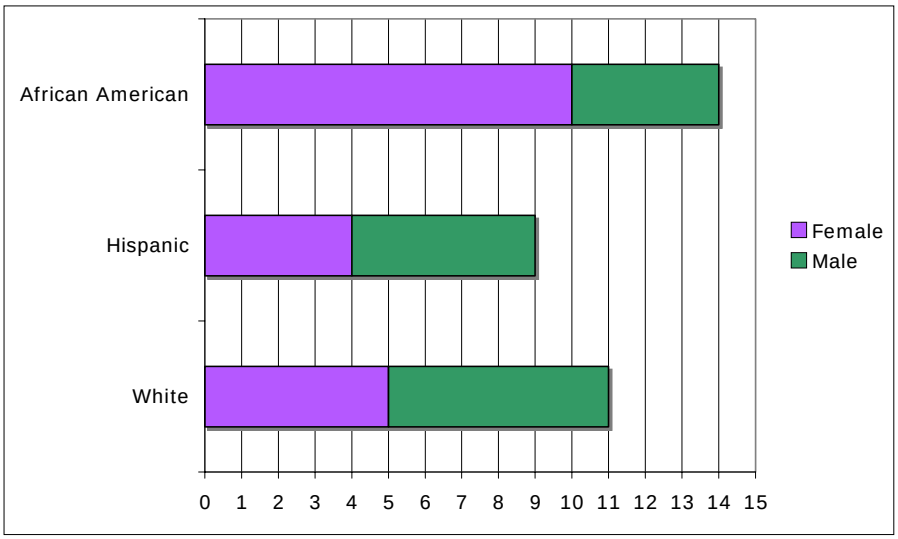


Figure 4.1: Student interview demographics

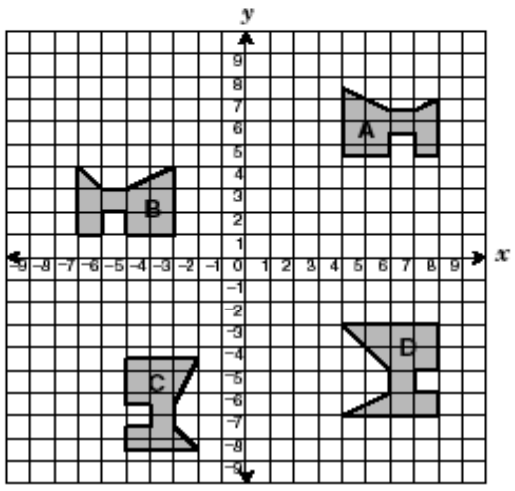
Each student worked on average around fourteen questions explaining his or her solution process as he or she worked. In order to get interview data on a broad range of items, the interviewees were not all given the same set of items. Each item included in this discussion was answered by at least thirteen interviewees. A more detailed discussion of the response data for selected items will follow later in this chapter.

Five content area specialists reviewed the TAKS items: two experienced high school teachers, two university-level Mathematics Education professors and one university-level Mathematics professor. The two high school teachers conducted their review as a team so only one survey was returned. The respondents were asked to review

each item and record the correct answer for the item, the primary objective covered by the item and the cognitive levels required to solve the item. Not all surveys were returned with responses for all of the items, however each item was reviewed by at least two content area specialists. A table with all survey results can be found in Appendix B. The survey results for the correct answer and primary objective were compared to the TEA stated answers and objectives for each item. TEA does not explicitly state the cognitive level needed to solve each item, so no comparison of this category was made; however, these data were useful in helping to interpret other results in this study.

In regards to the correct answer for each item, in general, the respondents agreed with the TEA stated correct answer. In sixteen of the items, however, the specialists did not agree on the correct answer. It is this author's opinion that these discrepancies are due to careless errors on the part of the respondents. For example, for item number 10, shown in the figure below, one respondent selected answer choice F, another selected answer choice H and two selected answer choice J.

10 Which pair of the following polygons is congruent?



F Polygon A and Polygon C
G Polygon B and Polygon D
H Polygon A and Polygon B
J Polygon B and Polygon C

The item asks which two figures are congruent and it is clear in this instance that only two of the figures are in fact congruent: figures B and C, which is answer choice J. Of course it is easy to see how a careless mistake could be made with this item. All of the figures have the same basic shape. Only by carefully examining the differences in lengths of the sides can one determine which two are in fact the same. One other item, number 59, also had three separate selections for the correct answer choice.

59 Which equation will produce the widest parabola when graphed?

A $y = 2x^2$

B $y = -6x^2$

C $y = -0.6x^2$

D $y = 0.2x^2$

In item #59, one respondent chose answer A, another chose answer B and two chose answer D. Again, this item is straightforward with little room for interpretation. When all four choices are graphed on the same set of axes, answer choice D clearly yields the widest parabola. This item was not, however, a particularly difficult item for the students; as a matter of fact, out of the thirteen students interviewed on this item, only one answered incorrectly.

In the remaining fourteen cases, only one respondent selected an answer different from the TEA-stated answer. These items were numbers 9, 13, 14, 19, 20, 21, 24, 27, 28, 37, 41, 45, 51 and 53. In each instance, it appears that the answer was selected due to a careless mistake on the part of the respondent, as was demonstrated in the case above. There appear to be no real discrepancies between the TEA-stated correct answer and the content area specialists' opinions, just many careless mistakes on the part of the content area specialists. The fact that the content area specialists were prone to careless mistakes does make one wonder how common careless mistakes are for students taking the exam.

Although verifying the answers for each item is important, the primary purpose of administering this survey was to get some sort of measure of what specialists in the field see as the primary objectives and cognitive levels the items on the test are measuring. The content area specialists were given a list of objectives to choose from for each item. The list paralleled the list of objectives published by TEA, with an additional category of "objective not found in list" added at the end.

(A) Functional Relationships

(B) Properties and Attributes of Functions

(C) Linear Functions

(D) Linear Equations and Inequalities

(E) Quadratic and Other Non-linear Functions

(F) Geometric Relationships and Spatial Reasoning

(G) 2-D and 3-D Representations

(H) Concepts and Uses of Measurement and Similarity

(I) Percents, Proportions, Probability and Statistics

(J) Mathematical Processes and Tools/Application of Mathematics

(Z) Objective not found in above list (please specify)

Although there were only a few items in which all of the content area specialists agreed unanimously with the TEA-stated objective, in many cases at least a majority of the content area specialists did (see the chart below).

Level of agreement with TEA	**Item Numbers**
High (3-4 CAS)	2, 7, 10, 14, 21, 35, 39, 40, 42, 53, 60
Majority (2-3 CAS)	1, 6, 8, 13, 17, 19, 24, 27, 31, 37, 45, 46, 48, 55, 59
Partial (1 CAS)	3, 4, 5, 12, 15, 18, 23, 28, 29, 32, 33 34, 38, 41, 43, 44, 50, 51, 54, 56, 58
No Agreement	9, 11, 16, 20, 22, 25, 26, 30, 36, 47, 49, 52, 57

The items I wish to focus on are the ones where there was no agreement between the TEA and the content area specialists. In some of the instances, the specialists agreed with one another, but their responses did not match the TEA-stated objectives, however in other instances, there was not agreement even within the specialists' responses.

The content area specialists were not given much detail in regards to the objective levels so as not to constrain their responses. The exact protocol given to the content area specialists can be found in Appendix A. Some of the discrepancies could be due to differences in interpretation of what exactly each objective entails. In the next section of this chapter, I will discuss selected items by combining the data collected through the content area specialists' surveys and the data collected through student interviews.

Item 49 - The Fertilizer Problem

49 The table below shows the cost of fertilizer, depending on the amount purchased.

Cost of Fertilizer

Number of Pounds	Cost
5	$1.95
20	$6.95
50	$15.95
100	$28.95

Which conclusion can be made based on information in the table?

A The cost of 10 pounds of fertilizer would be more than $4.00.

B The cost of 200 pounds of fertilizer would be less than $57.00.

C The cost of fertilizer is always more than $0.35 per pound.

D The cost of fertilizer is always less than $0.30 per pound.

I was especially interested in this problem because in my initial review of the test, I did not believe any of the listed answer choices to be correct. Three of the choices can be ruled out mathematically, however the fourth cannot be confirmed from the

information given. It could be argued that this is therefore the best answer because by default it is correct; however, I was interested to see what others would say about the matter. None of the content area specialists flagged this item as being out of the ordinary and all agreed with the TEA-stated answer choice. There was, however, a discrepancy between the TEA-stated objective and the opinion of the majority of the content area specialists. Three out of four respondents listed the primary objective for this item as "Functional Relationships" but TEA stated the objective for this item as "Percents, proportions, probability and statistics." What is interesting is the sub-objective listed for this item. According to TEA, this item is supposed to assess if students can "recognize the misuses of graphical or numerical information and evaluate predictions and conclusions based on data analysis" (TEA, 2004d, p. 6). Although this problem does appear to be an illustration of misuses of numerical information, it is not clear how a student would demonstrate competence with this objective in solving this problem.

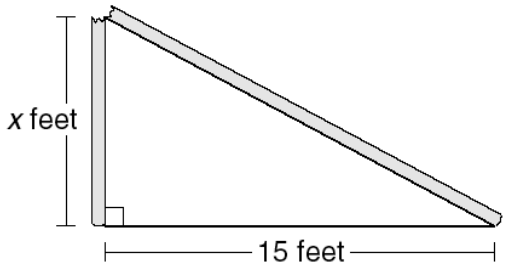
Since I was particularly interested in this item, every student interviewed was given this item to work through. The most common solution strategy for this item was to somehow find price per pound and then try to eliminate answer choices. Only two students tried to enter the data into the graphing calculators in order to get a scatter plot. Out of the 34 students interviewed on this item, fourteen answered B, the answer stated by TEA as the correct answer. Of these fourteen, four students eliminated the other three choices and selected B without verifying whether or not it could be true. Two students reasoned out that B could be true because the price goes down the more you buy. Two other students chose B because it was the only one that could be true. One student eliminated all four choices and then selected B because it was the closest to being true. The other five students chose B based on what I will call non-mathematical reasoning: "B just seems right," etc.

In addition to the one student who selected B after eliminating all of the choices, only two other students eliminated all of the choices and then had to decide which answer was best. The most common mistake with this problem was to assume the price per pound was constant in which case answer C is the best answer. Nine students solved the

problem this way. From the student interview data, it does not appear the students were using data analysis to solve this problem but rather a general test-taking strategy of elimination of choices until the "correct" one is found. Therefore, the item does not appear to be measuring the TEA-stated objective.

Item 22 – The Broken Pole

22 A wooden pole was broken during a windstorm. Before it broke, the total height of the pole above the ground was 25 feet. After it broke, the top of the pole touched the ground 15 feet from the base.



How tall was the part of the pole that was left standing?

F 8 ft

G 10 ft

H 17 ft

J 20 ft

This is another example of an item where none of the content area specialists agreed with the TEA-stated objective. This question asks students to find the height of what is left standing of a broken telephone pole given the initial height of the unbroken pole and the distance the top of the broken pole is away from the base. Three out of four of the content area specialists stated the primary objective for this item as Geometric Relationships and Spatial Reasoning. When I looked at this item, I would have also placed it in that category, assuming it was intended to see whether students could apply the Pythagorean theorem. Although this problem can be solved with a direct application

of the Pythagorean theorem, TEA intended this problem to asses the problem solving skills of students, specifically students' abilities to "use a problem solving model that incorporates understanding the problem, making a plan, carrying out the plan, and evaluating the solution for reasonableness" (TEA, 2004d, p. 6).

Upon closer examination of the problem, one can see that a student can reason out the correct answer by eliminating the answer choices rather than applying the Pythagorean theorem. Out of the 22 students who answered this question, thirteen answered correctly. Eight students used the Pythagorean theorem but only to substitute in answer choices. Three students physically measured the missing side and two others reasoned out that F was the only possible answer because the side would have to be less than ten. Of the eight students who answered incorrectly, four merely subtracted fifteen from twenty-five and got an answer of ten. Two others tried to measure/estimate the length of the side. The other two tried to use the Pythagorean theorem, but set it up incorrectly. Below, I include an excerpt from an interview transcript illustrating how a student could reason out the correct answer.

Respondent	hmm…I think I will just plug in the answers.
Interviewer	Oh, okay. Plug them in to do what?
Respondent	I going to do 8 squared plus 15 squared inside a square root to se if it comes out to 25.
Interviewer	Ok. Alright.
Respondent	And that comes out to 17 so that's not it. 10 squared plus 15 squared comes out to 18, so that's not it. 17 squared plus 15 squared…maybe I'm doing this wrong. Because that is not going to be 25 so I'm not looking for 25. I'm looking for them to add up to 25.
Interviewer	oh, Okay
Respondent	hmm…so…maybe it was 8.
Interviewer	How come?

Respondent 8 plus 17 equals 25. Yep, because the whole pole is 25. So side x and the long side of the triangle have to add up to 25.

In general, students did use some sort of problem solving method with this problem rather than directly using the Pythagorean theorem. In this case, the interview data support the TEA-stated objective rather than the content area specialist reviews.

Item 26 - Surface area of a cube

26 If the surface area of a cube is increased by a factor of 4, what is the change in the length of the sides of the cube?

F The length is 2 times the original length.

G The length is 4 times the original length.

H The length is 6 times the original length.

J The length is 8 times the original length.

Students are asked to find what effect changing the surface area of a cube has on the length of the sides of the cube. In this case, all three content area specialists who reviewed this item selected Geometric Relationships and Spatial Reasoning as the primary objective, however TEA lists Measurement as the primary objective. Here, I believe the difference is due to the way TEA defined the sub-objectives for the overarching measurement objective, which the content area specialists did not have access to. The specific sub-objective listed for this item states that students should be able to "describe the effect on perimeter, area, and volume when length, width, or height of a three-dimensional solid is changed and applies this idea in solving problems" (TEA, 2004d, p. 5). This sub-objective defines a very specific topic that is seems to be covered with this question.

Once again, the only way to know for sure what this item is actually measuring is to look at the student interview data. Out of the 24 students who answered this question, sixteen answered correctly, however only ten of those sixteen had what I will call an

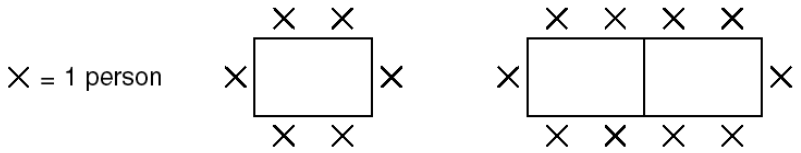
appropriate mathematical solution. Twelve students attempted to set up an equation with example numbers while the other twelve took other various approaches. Of the twelve who tried to set up an equation, only five were able to do so correctly. The others had difficulty in knowing what "by a factor of four" meant. Some tried to add four and others tried to raise the number to the fourth power. The students who simply added four did not get an answer that matched any of the choices, however F was the closest, so they chose F. The really interesting solutions came from the three students who interpreted the term "factor" in another way. These students looked at the questions and reasoned that if we increase by a factor of four, two is the only number listed that is a factor of four, so the answer must be F. The excerpt below shows this.

Respondent	Ok. If the surface area of a cube is increased by a factor of 4, what is the change in the total length of the sides. Of, so surface area for a cube is, um…its equals 6 s squared and it says the surface area of a cube is increased by a factor of four what is the change in the length of the sides of the cube. So I know that a factor of four is 2 so that would just be my answer the length is 2 times the original length.
Interviewer	Because?
Respondent	Uh, 2 is a factor of four.
Interviewer	2 is a factor of four, ok.
Respondent	And it says it is increased by a factor of four. So that just seems like the right answer.

This sample size is too small to know whether this was a common belief, however it does warrant further investigation. For many of the students interviewed, this item did in fact seem to be testing the TEA objective, however it is still possible for a student to know little about what the problem is actually asking and come up with the correct answer.

Item 36 – Banquet Tables

36 For a sports banquet Coach Mackey must use the rectangular tables in the school cafeteria. The diagram below shows the seating arrangements that Coach Mackey can use at 1 and 2 tables.



Which expression can be used to determine the number of people who can sit as a group if y tables are joined to form 1 long table?

F $6y$

G $4(y + 1)$

H $3(y + 1)$

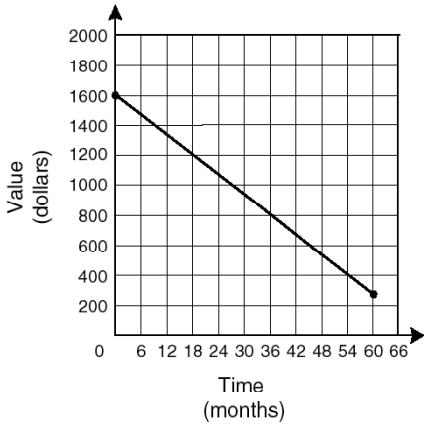
J $2(2y + 1)$

In this question, students are asked to find an expression that represents the number of people that could sit at a banquet table. Here, two of the three of specialists surveyed on this item listed Functional Relationships as the primary objective. The TEA-stated objective for this item is Mathematical Processes and Tools where the student is expected to demonstrate an ability to "make conjectures from patterns or sets of examples and non examples" (TEA, 2004d, p. 6). Again, I believe the difference could be a result of the limited information the content area experts were given on the specific sub-objectives. This item could be used to assess student understanding of functional relationships, i.e. the relationship of the number of people to the number of tables, however, TEA intended the item to measure a different objective.

Fourteen students were interviewed on this item, nine of whom answered correctly. Of these nine, only two recognized a pattern and were able to pick out the correct relationship. All of the other seven substituted numbers into each relation given until they found one that worked for both examples given. For these seven students, the item was not measuring the TEA-stated objective, but rather a general test taking strategy of working backwards.

Item 47 – Computer depreciation

47 The graph below shows the decrease in the value of a personal computer over a period of 60 months.



Which is a reasonable conclusion about the value of this personal computer during the time shown on the graph?

A Its value at 18 months was twice its value at 36 months.

B Its value at 36 months was half its value at 54 months.

C It depreciated $200 every 12 months.

D It depreciated $400 every 18 months.

Here, students were asked to look at a graph of the value of a personal computer over a period of 60 months and select a reasonable conclusion that could be drawn. Two of the three content area specialists surveyed on this item selected Linear Functions as the primary objective, while the third selected 2-D and 3-D Representations. The TEA-stated objective for this item is Properties and Attributes of Functions. All three of these objectives are similar in nature, so it is easy to see how each one could apply in this case. The TEA-stated sub-objective is that a student will be able to interpret "situations in terms of given graphs" (TEA, 2004d, p. 1), which seems to be the case for this item.

For this item, thirteen students were interviewed. All but one of the thirteen tried to interpret the graph in order to answer the question. The one who did not stated that she did not know what the word "depreciated" meant, which was used in two of the answer choices, so she guessed between the two answers she could understand. Of the others,

nine students answered correctly. Each one of the nine went through all of the answer choices comparing the answer choice to the data shown on the graph until they found one that was true. It appears, at least from the small sample here, that the students were demonstrating their ability on the TEA-stated sub-objective.

Item 9 - Cubes in a box

9 How many 2-inch cubes can be placed completely inside a box that is 8 inches long, 2 inches wide, and 6 inches tall?

A 8

B 12

C 24

D 48

In this item, students were asked to determine the number of cubes that could be placed in a box with given dimensions. Again, none of the content area specialists selected the TEA-stated objective, Mathematical Processes and Tools, as the primary objective. Instead, the majority selected Geometric Relationships and Spatial Reasoning. While this problem does involve spatial reasoning, TEA intended this problem to measure students' problem solving skills. The sub-objective listed for this problem states that students will "select or develop an appropriate problem solving strategy from a variety of different types, including drawing a picture, looking for a pattern, systematic guessing and checking, acting it out, making a table, working a simpler problem, or working backwards" (TEA, 2004d, p. 6).

In the 23 interviews conducted on this item, nearly all of the students calculated the volume of the box first and then tried to see how many cubes would fit. Not all students took into account the volume of the cubes and simply divided the volume of the box by two. The wording of the question may have thrown some of the students off since it says the cubes are 2-inches. Some students may have interpreted that to mean the cubes were two inches in volume rather than two inches on a side. Only thirteen of the 23 students interviewed selected the correct answer choice. Of these, seven calculated the

volume of the box and divided it by the volume of the cubes. Five others drew three-dimensional sketches in order to see how many cubes would fit. The other student who answered correctly only did so because he remembered working through this problem in class and knew what the answer was supposed to be.

The students who drew a picture to solve this problem were using a problem solving process as described in the above TEA sub-objective, however, it is unclear whether the students who calculated the volume were using a problem solving process or just going through a mathematical process for calculating volume.

For the next few items discussed, no student interview data were available; therefore, I will only discuss the content area specialists' responses in comparison with the TEA-stated objectives. In each case, as with the items discussed above, there was no agreement between the primary objectives stated by the content area specialists and the ones stated by TEA. It is this author's opinion that these differences are most likely due to way TEA defines sub-objectives for each of the primary objectives.

Item 11 – Finding average speed

11 Let a represent the average speed in miles per hour a car traveled on a trip. Let $f(t)$ represent the distance in miles the car had traveled t hours after the beginning of the trip. The function $f(t)$ is best represented by —

A $t^2 + a$

B at^2

C $t + a$

D at

Students are asked to find a relationship that represents the distance a car has traveled in relation to its average speed and time on the road. For this item, the majority of the content area specialists selected Functional Relationships as the primary objective. The TEA-stated objective is Properties and Attributes of Functions. These two objectives are closely related; therefore, the discrepancy between the content area specialists and TEA is understandable. If we look at how TEA defines the sub-objective we see better

where this item fits in. TEA defines the sub-objective for this item as "the student uses symbols to represent unknowns and variables" (TEA, 2004d, p. 1).

Item 16 – Parallel lines

16 Which of the following best describes the graph of the equations below?

$$2y = 3x + 2$$

$$4y = 6x + 1$$

F The lines have the same y-intercept.

G The lines have the same x-intercept.

H The lines are perpendicular.

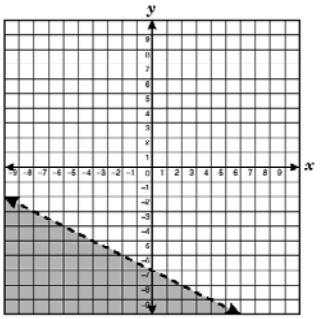
J The lines are parallel.

Here, the content area specialists split between two objectives: Properties and Attributes of Functions and Linear Equations and Inequalities. This problem asks a student to compare two lines and state what relationship the two lines have to one another. It is easy to see how the content area specialists chose the objectives they did. However, TEA considers this item to fall under the 2-D and 3-D representations objective, defining the sub-objective for this item to be dimensionality and geometry of location. Under this sub-objective, students are expected to use "slopes and equations of lines to investigate geometric relationships, including parallel lines, perpendicular lines and (special segments of) triangles and other polygons" (TEA, 2004d, p. 4). Here again, one can guess that if the content area specialists had been given this level of detail, the responses might have been different.

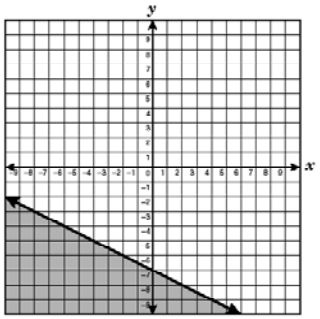
Item 20 – Graph of an Inequality

20 Which graph best represents the inequality $x + 2y \leq -14$?

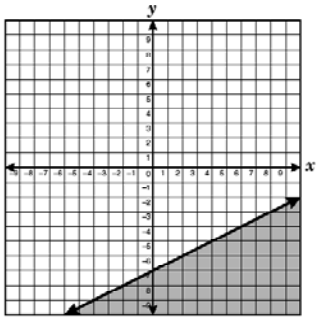
F



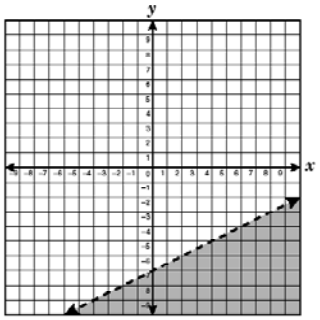
H



G



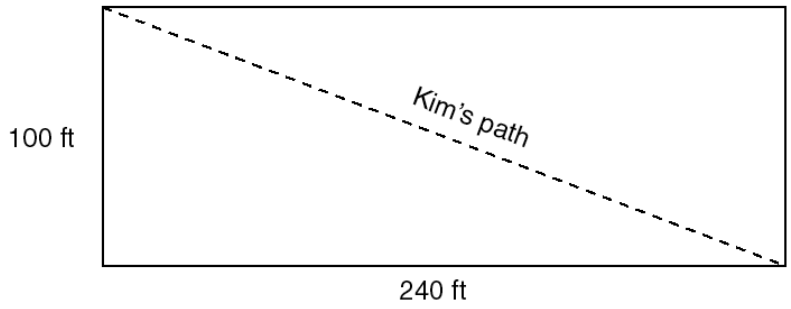
J



This question asks students to identify the graph of a given inequality; therefore, it is not unexpected that the content area specialists would unanimously select Linear Equations and Inequalities as the objective for this item. Again, the way TEA has defined the objectives makes a difference in the way this item would be interpreted. TEA lists the primary objective of this item to be Functional Relationships. Graphs of inequalities are explicitly named in the sub-objective where it is listed that students should be able to represent "relationships among quantities using [concrete] models, tables, graphs, diagrams, verbal descriptions, equations, and inequalities" (TEA, 2004d, p. 1).

Item 25 – Walking a diagonal Path

25 Kim walked diagonally across a rectangular field that measured 100 feet by 240 feet.



Which expression could be used to determine how far Kim walked?

A $2(100 + 240)$

B $\sqrt{100} + \sqrt{240}$

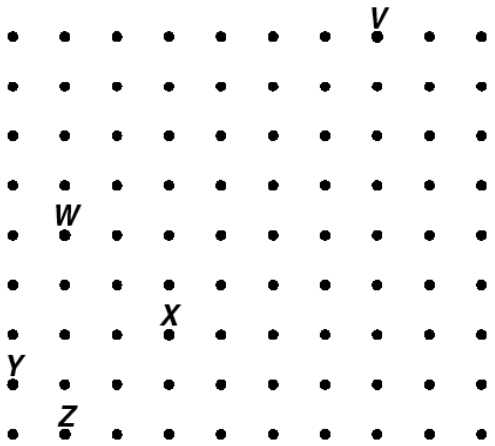
C $\frac{100 \times 240}{2}$

D $\sqrt{(100^2) + (240^2)}$

This item asks students to find the length of a path that cuts diagonally across a rectangular field. Here again, the majority of content area specialists surveyed on this item selected Geometric Relationships and Spatial Reasoning as the primary objective for this item. As with item 22 discussed above, this problem can be solved through direct application of the Pythagorean Theorem, but TEA lists use of Pythagorean Theorem under the Measurement objective rather than under the Geometric Relationships objective. This is another case of differences in interpretation of what each objective consists of.

Item 30 – Finding slope on a Geoboard

30 As part of a classroom assignment, Kimberly was given this geoboard to model the slope of $\frac{2}{3}$.



If the peg in the lower left-hand corner represents the origin on a coordinate plane, where could Kimberly place a rubber band to represent the given slope?

F From peg *V* to peg *W*

G From peg *V* to peg *X*

H From peg *V* to peg *Y*

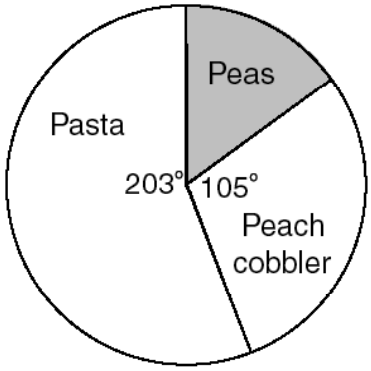
J From peg *V* to peg *Z*

Here students are asked to determine which two, labeled points on a Geoboard grid represent a slope of two-thirds. In this item, there was no real consensus amongst the content are specialists as to what the primary objective was. Two surveys stated Geometric Relationship and Spatial Reasoning as the primary objective, one stated Concepts and Uses of Measurement and Similarity and the other stated Linear Functions. According to TEA, this item falls under the Mathematical Processes and Tools objective and was intended to test whether students could "communicate mathematical ideas using

language, efficient tools, appropriate units, and graphical, numerical, physical, or algebraic mathematical models" (TEA, 2004d, p. 6).

Item 52 – Arc Length

52 A frozen dinner is divided into 3 sections on a circular plate with a 12-inch diameter.



What is the approximate length of the arc of the section containing peas?

F 3 in.

G 21 in.

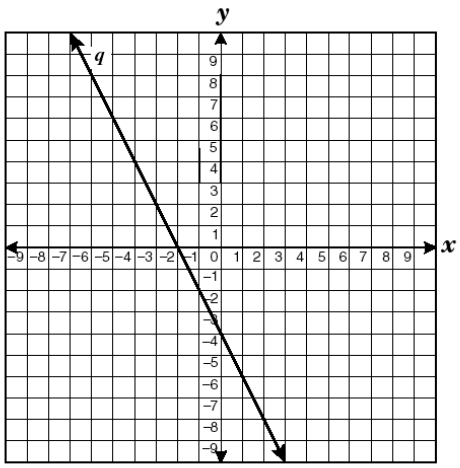
H 16 in.

J 5 in.

In this item, students are given a circle divided into three unequal pieces. Angles for two of the three pieces are given and students must calculate the arc length that corresponds to the missing angle. As with item 30 above, there was no real consensus amongst the specialists as to what the primary objective was; two selected Geometric Relationships and Spatial Reasoning, one chose Linear Equations and Inequalities and the other chose Mathematical Processes and Tools. TEA states the primary objective as Measurement, however, again the sub-objective under Measurement specifically lists arc length. According to TEA, the students should be able to find "areas of sectors and arc lengths of circles using proportional reasoning" (TEA, 2004d, p. 5).

Item 57 – Equation of a line parallel

57 Line q is shown below.



Which equation best represents a line parallel to line q?

A $y = -\frac{1}{2}x + 4$

B $y = \frac{1}{2}x - 3$

C $y = 2x - 5$

D $y = -2x + 1$

Here students are asked to find the equation of a line parallel to the one shown. Two of the three content area specialists who reviewed this item stated that the primary objective was Linear Equations and Inequalities. TEA states that the primary objective is instead 2-D and 3-D representations. Again, the difference is most likely due to the fact that TEA explicitly states that relationships between parallel lines falls under this objective where the students will use "slopes and equations of lines to investigate geometric relationships, including parallel lines" (TEA, 2004d, p. 4).

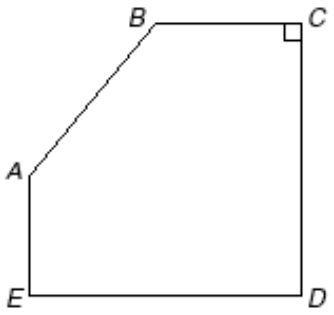
In the cases listed above, the objectives selected by the content area specialists seem reasonable with the level of detail given to them about each objective. The differences in the perceived primary objective for each item may not be true differences but rather artifacts created by the way TEA defines its objectives and sub-objectives. Further interviews with students and these, as well as other, content area specialists would be necessary in order to determine if this is in fact the case.

To conclude this portion of the chapter, I will share some other interesting results that came from the student interviews. In each of the following cases, there was at least some agreement between the content area specialists surveyed and TEA as to the primary objective for each item, however the student interviews provide a different perspective.

<u>The non-multiple-choice item</u>

Item 21 – Find the Missing Angle

21 In the figure shown below, $\overline{BC}$ is parallel to $\overline{ED}$, and $\overline{AE}$ is perpendicular to $\overline{ED}$. The measure of $\angle ABC$ is 130°.



What is the measure of $\angle BAE$ in degrees?

Record your answer and fill in the bubbles on your answer document. Be sure to use the correct place value.

This is the one non-multiple-choice item on the test. Here students have to "bubble in" the correct response on their answer sheets. This is a very interesting problem because a student is unlikely to simply guess the correct answer. Out of the twenty-seven students who were interviewed on this question, only nine answered correctly. Another

nine students stated the angle was the same as the one given, so the answer was 130 degrees. The remaining nine students came up with various other answers. No students interviewed simply guessed the correct answer. All nine students who answered correctly did so by applying geometric knowledge about the figure to find the missing angle. (see the transcript excerpt below)

Respondent	Uh, lets see. Its 30 degrees that has to be 180 it's got to be a flat line so that's got to be 50 degrees. And that's got to be 90.
Interviewer	So show me what you're doing there.
Respondent	Alright. So in the triangle here, it's going to be a right triangle because it is part of a square so that has to be 90 degrees.
Interviewer	Ok
Respondent	And I've learned that when um…you have this angle right here, and I can't draw, this angle has to be, whatever this angle is if you move in a straight line, its 180 degrees. but we already know that part of the angle is 130 so we know that this angle is has to be 50 degrees to equal 180.
Interviewer	Ok
Respondent	We've got to find this one but we know this one, so if that one's 90 degrees and that one's 50 degrees this one has to be 40 degrees. They both have to equal up to 90.
Interviewer	Ok
Respondent	So if this one's 40, we can use the same principle and find out that this has to be 140. So I would write in the answer 140.
Interviewer	Ok

This is one of the rare instances where all four content area specialists agreed not only with one another but also with the TEA-stated objective of Geometric Relationships

and Spatial Reasoning. The TEA-stated sub-objective for this item is stated as follows, "[students will use] numeric and geometric patterns to make generalizations about geometric properties, including properties of polygons, ratios in similar figures and solids, and angle relationships in polygons and circles" (TEA, 2004d, p. 4). Judging by the student interview responses for this item, this item seems to actually measure the objective it was designed to measure.

Getting the right answer with an incorrect solution.

Item 2 Simple Probability

2 In a high school auditorium, 1 junior and 2 sophomores are seated randomly together in a row. What is the probability that the 2 sophomores are seated next to each other?

F $\frac{1}{9}$

G $\frac{1}{3}$

H $\frac{2}{3}$

J $\frac{5}{6}$

Students are asked to find the probability that, given one junior and two sophomores, the two sophomores will be seated next to one another. This question seemed straightforward. In order to solve the problem, one can enumerate the possibilities and count up how many have two sophomores sitting next to one another. The content area specialists all agreed with the TEA-stated objective: Percents, Proportions, Probability and Statistics. Specifically, TEA says this item will measure a students' ability to "find the probabilities of compound events" (TEA, 2004d, p. 5). Out of the 23 students interviewed on this item, twenty answered correctly, however not all of the twenty students exhibited valid mathematical reasoning. Ten students enumerated the possibilities and counted up, however the other ten simply stated that since two out of the

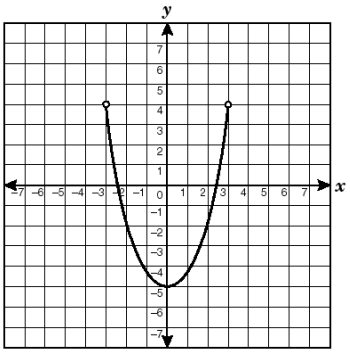
three were sophomores, the probability would be two-thirds. (see transcript excerpt below)

Interviewer	So tell me how you got that. Walk me through it
Respondent	Well this one, its probably wrong, but looking at it there is a lot of sophomores. And there is only, there's two sophomores and only 1 junior, so I'm just going to put 2 out of three, but that could be wrong. It could be tricking me with the item. It could be like J because it seems like it would be allowed to, that it would happen a lot. But I guess I'm going to stick with 2 out of 3.
Interviewer	Ok.

While in this particular case this is true, it is not the general rule. If the question had asked instead about 3 out of 5 students, the answer is not 3/5. This item does not appear to be truly measuring what it is intended to measure and may even foster a misconception about this type of probability.

Item 27 Finding domain of a function

27 What is the domain of the function shown on the graph?



A $-3 \leq x \leq 3$
B $-3 < x < 3$
C $-5 < x \leq 4$
D $-5 \leq x < 4$

This question asks students to state the domain of a function given its graph. Three out of four content area specialists agreed with the TEA-stated objective of Properties and Attributes of Functions. Out of the fifteen students who answered this question, thirteen answered correctly. Several interesting things happened with this problem. First, both of the students who answered incorrectly, stated the correct definition of domain, the x-values a function can take on, however both chose A because since the line is solid, the answer must be 'less than or equals to'. Of the thirteen who answered correctly, only six explicitly stated knowledge of domain. Six students selected choice B because they knew that the open circles meant strictly less than and choice B is the only one with two strictly less than symbols. The others all have a less than or equal to somewhere. (See transcript excerpt below).

Interviewer	Tell me why you're crossing those out.
Respondent	Well these are easy, 'cause you can look at them and you know that you see right here it is an open circle on both of them, so you automatically know that it's not going to equal it. Every single one of those has an equals except for B.
Interviewer	Ok

Again, the answer is technically correct, however the students did not exhibit specific knowledge of domain.

Vocabulary Issues

Item 19 Sewing sequins

19 Megan is using an equilateral triangle as part of a design on a sweatshirt. Each side of the triangle is 12 inches long. Megan is sewing a line of sequins from the midpoint of one side of this triangle to the opposite vertex. Approximately how long will the line of sequins be?

A 13.4 in.

B 10.4 in.

C 8.5 in.

D 5.2 in.

Out of the thirteen students interviewed on this item, eight were able to answer correctly. What is interesting about this problem is the fact that two students stated that they did not know what sequins were. In one case, the student was able to just ignore the word, move on and correctly solved the problem, however in the other case, the student stated that since he did not know what sequins were, he could not solve the problem and guessed. He did manage to guess the correct answer, but it was just happenstance. Although both students were able to select the correct answer, this is still an issue for further investigation.

Use of Calculator

Although the students interviewed used the graphing calculators to varying degrees, two problems were consistently solved solely with the graphing calculator.

Item 59 Widest Parabola

59 Which equation will produce the widest parabola when graphed?

A $y = 2x^2$

B $y = -6x^2$

C $y = -0.6x^2$

D $y = 0.2x^2$

In this question, students are asked to state which equation will produce the widest parabola when graphed. All of the content area specialists agreed with the TEA-stated objective: Quadratic and Other Non-Linear Functions. Out of the thirteen students who answered this item, only two stated any prior knowledge of the effect of a constant multiplier to a power function. In one case, the student was sure of his answer and did not use the calculator to verify. He stated that larger numbers make parabolas wider, so choice A is correct. It is unclear if in a live testing administration this student would answer the same way without using the calculator to check. The other student who stated knowledge of the effect of the multiplier remembered correctly that a smaller number will make the parabola wider, however he used the calculator to verify his answer. Out of the other eleven students who answered correctly, none stated any knowledge of what the multiplier would do. Each of the eleven simply graphed the four equations and chose the widest one. The excerpt below shows what one student said about this problem.

Respondent	Which equation will produce the widest parabola when graphed? Alright, again, this is one of my favorites. It is just a gimme question.

Interviewer	Yeah? How come?
Respondent	In dealing with the graphing calculator, since we've, I only really learned how to do it with a regular calculator and sometimes without it, but now when they give us now graph the equation, well its like, well we already have all of these. All I have to do is plug them in. I mean I guess maybe they give these questions so you don't (mumbles) maybe its these calculator questions where they go oh well let's see if they know how to use a calculator or it depends on what they're thinking. I don't know. It's easy for me. It's easy for any one to just plug it in.

It is questionable how much actual algebra knowledge was involved in answering this question. From the students interviewed, it appears that this item tests calculator use more than anything else. Recall from the earlier results from the content area specialists that this was one of the most frequently missed items for the specialists. This could be due to the fact that they might not have used a graphing calculator when working through the problems. This problem becomes much easier when solved with a graphing calculator.

Item 42 - Find the roots

42 Which ordered pair represents one of the roots of the function $f(x) = 2x^2 + 3x - 20$?

F $(-\frac{5}{2}, 0)$

G $(-4, 0)$

H $(-5, 0)$

J $(-20, 0)$

In this question, students were asked to find one of the roots to a given quadratic function. Again, all of the content area specialists who responded to this item listed the same objective as the one stated by TEA. Again, this was listed as Quadratic and Other Non-Linear Functions. Of the thirteen students who answered this question, all chose the correct answer choice. All but three used the calculator to graph the function and find the roots. Two of the three chose G because it was one of two whole-number answer choices. The third student plugged each of the values into the function to find the one that yielded a zero. The other ten used the graphing calculator in various ways such as reading the root straight from the graph, reading the root from the table or using the TRACE function. Here again, this problem seems to be testing use of the calculator much more than knowledge of algebra. This problem does require more knowledge than item #59, listed above, because a student has to know what a root is, however, all of the choices have a zero as the y-coordinate and the answer G is the only one that falls on the function so there is no way to know if students are finding a root or just a point that happens to be on the line.

General Comments About Results

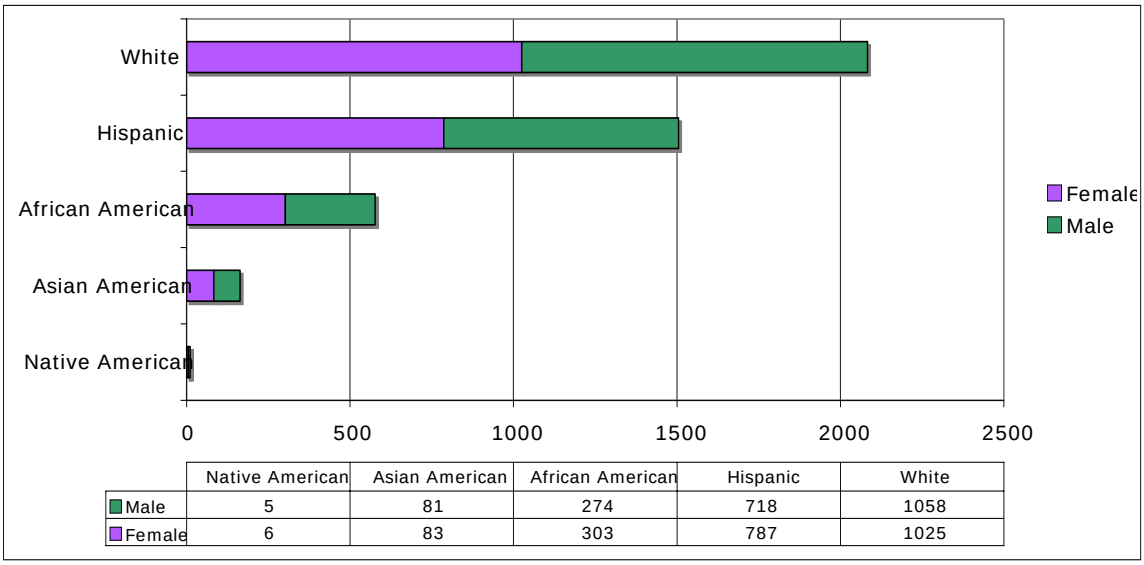
Both of these measures are qualitative and are therefore subjective in nature. In examining the validity of an assessment it is necessary to have various forms of evidence. The content area specialist surveys provided valuable insight; however, the most useful insight came from the student interview results. No matter what objectives the test was designed to measure or what specialists in the field believe the test measures, it is really how the test-takers handle the items that counts toward the actual objective measured. In general, I found that the content objectives for each item were not as important as 1) students test-taking skills and strategies, and 2) student use of technology, specifically use of a graphing calculator.

Data Set from TEA

The data set provided by TEA was a random sample of 5,000 students who took the Exit Level TAKS mathematics portion in the spring of 2004. Demographic data given for each student consisted of the region, district and campus the student belonged to, gender, ethnicity, and Limited English Proficient (LEP) status for each student as well as whether the student was classified as economically disadvantaged. In addition, for each of the sixty items on the math portion of the test, each student's response was listed along with his or her total score and whether or not he or she passed the math portion of the exam. It is the policy of TEA to mask student scores if, at a particular campus, the student is a member of an ethnic group with less than twenty members. Therefore, out of the 5,000 students, item response data and total score data was visible for 4,811 students. Additionally, according to the item response data, 471 students left the entire test blank. It is unclear as to whether these students failed to record their answers appropriately or if in fact they did not answer any of the questions. Since this determination could not be made, these students were omitted from analysis. Therefore all analysis was conducted on a total of 4,340 students.

Demographic make-up of sample

The following charts and tables illustrate the distribution by gender and ethnicity of the sample of 4,340 students. Figure 4.2 below shows the breakdown of both gender and ethnicity for the sample of 4,340 students.



	Native American	Asian American	African American	Hispanic	White
Male	5	81	274	718	1058
Female	6	83	303	787	1025

Figure 4.2: Gender and ethnicity of sample

Just to give a frame of reference for how this compares to the state demographic data, I have included the tables below. According to TEA, in the spring of 2004, a total of 216,083 students were administered the math portion of TAKS. Of these students 110,533 were female and 105,371 were male (179 students did not provide gender data) (TEA, 2004e). As we can see in table 4.1, the proportions of males and females in the sample are comparable to the proportions for the state data.

Table 4.1: Gender proportions

Gender	SAMPLE	POPULATION
Female	50.78%	51.20%
Male	49.22%	48.80%

Table 4.2 shows the proportions of students for each ethnic subgroup in the sample as compared to the state population data. According to TEA, during this test administration, there were 105,149 White students, 74,238 Hispanic students, 27,873 African American students, 7,721 Asian students and 651 Native American students who

took the test (451 students did not provide any ethnic subgroup data) (TEA, 2004e). Again, the proportions in the population of all students who took the math portion of the spring 2004 TAKS and the proportions in my sample are comparable.

Table 4.2: Ethnic proportions

	SAMPLE	POPULATION
White	48.0%	48.8%
Hispanic	34.7%	34.4%
African American	13.3%	12.9%
Asian	3.8%	3.6%
Native American	0.3%	0.3%

Each year TEA sets the passing standard for each of the four content area TAKS assessments. For 2004, the passing standard for the mathematics portion was set at 24 items. This means that in order to pass the math portion, a student must get 24 or more items correct out of 60. Table 4.3 below shows for my sample the proportions of students, broken down by ethnicity, who did not meet the TEA-set standard for math and also in the population.

Table 4.3: Failure to meet TEA standard

	Did not meet standard - Sample %	Did not meet standard - Population %	Met standard - Sample %	Met Standard - Population %
Native American	18.2%	12%	81.8%	88%
Asian	1.2%	5%	98.8%	95%
African American	28.4%	27%	71.6%	73%
Hispanic	21.0%	22%	79.0%	78%
White	8.1%	9%	91.9%	91%

In order to better illustrate how these proportions compare to one another in my sample, I have also included figure 4.3 below. This chart shows the total proportion of each ethnic subgroup that failed to meet the standard, sub-divided by gender. The bar height depicts the total percentage of students who failed to meet the passing standard for each sub-group. The yellow portion of the bar represents the proportion of the female students in the subgroup that failed to meet the standard and the green bar represents the proportion of male students.

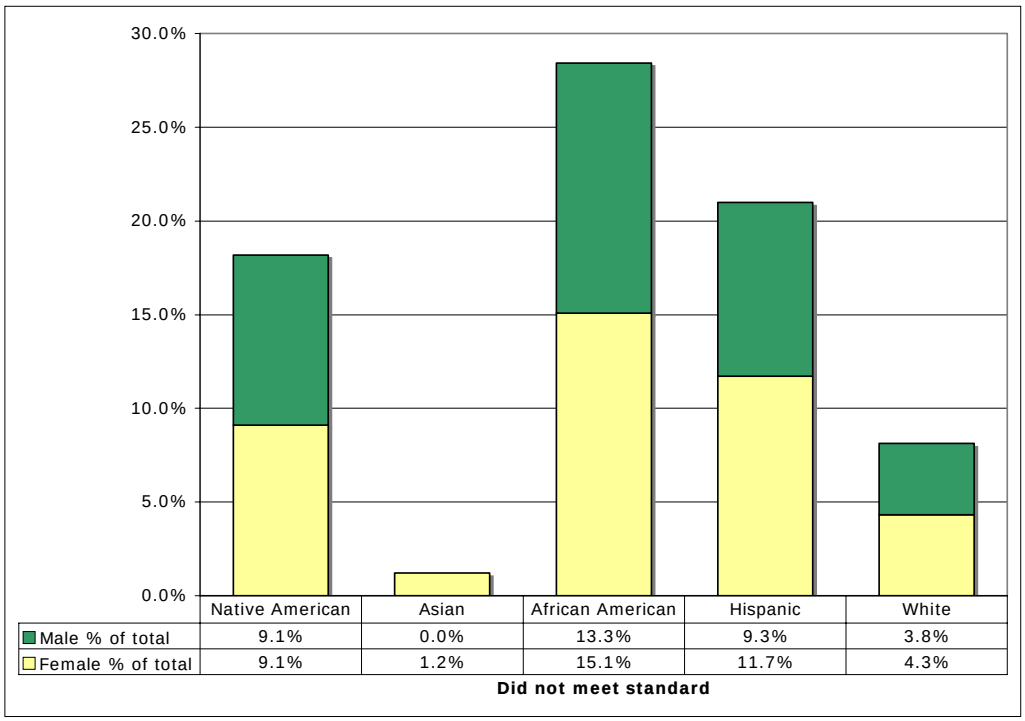


Figure 4.3: Proportion of ethnic subgroups that failed to meet TEA standard

We can see from this above chart that overall, African American students failed to meet the passing standard in the largest proportions, a total of 28.4% with Hispanic students next at 21%. This is in comparison to the 8.1% of white students and 1.2% of Asian students who failed to meet the standard. This difference in passing rates should be investigated further to determine the underlying causes. The Differential Item

Functioning analysis discussed later in this chapter looks into the performance differences item by item, but more work should be done with the test a whole.

The next chart is presented in order to give an idea about the overall difficulty of the items on the test. The chart lists ranges of p-values (proportions of students getting an item correct). Within each bar I have listed the actual item numbers falling into each category (see figure 4.4 below)

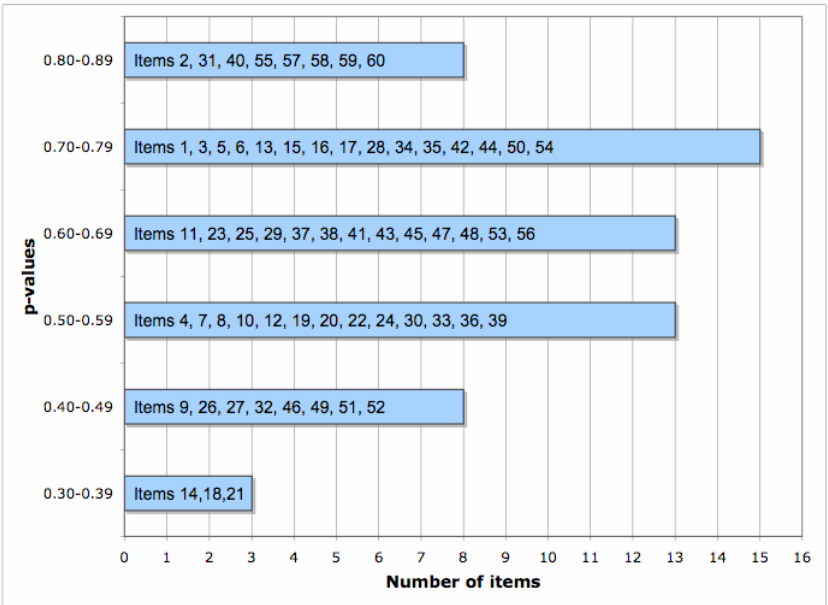


Figure 4.4: Items listed by P-Value

The figure above shows that of the 60 items on the exam, 23 items have p-values greater or equal to 0.70, (note this is one item less than the number students had to have right in order to pass). The item with the lowest p-value is the non-multiple choice item, item 21 discussed in the interview response section.

The preceding charts were shown and discussed in order to given the reader some insight into the general nature of the dataset. Two main analyses were performed on the data: factor analysis and differential item functioning. I will discuss the results for each in turn in the sections that follow.

Factor Analysis Results

In order to conduct a factor analysis on this data, first the item response data had to be converted to dichotomous, right-wrong data. For all responses, a “0”was assigned to incorrect responses and a “1” to correct responses. I then used SAS to create a tetrachoric correlation matrix, the correlation matrix used to analyze dichotomous data, from which factor analysis could be conducted. A table of the initial eigenvalues is given in Appendix C. Since each of the 60 items is considered a variable, the factor analysis yields 60 eigenvalues. Figure 4.5 below is the scree plot of the initial eigenvalues.

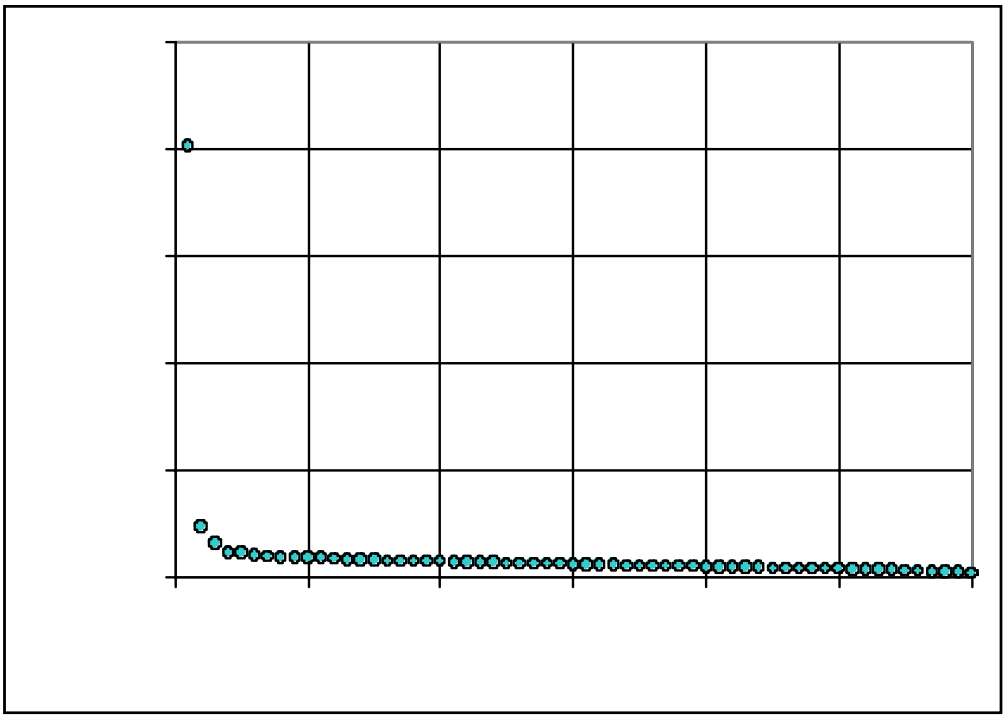


Figure 4.5: Scree plot of initial eigenvalues

After looking at the scree plot of the eigenvalues, I decided to retain three factors. As stated in Chapter 3, Methodology, the standard is to retain the factors above the “elbow” or bend in the scree plot. The first factor clearly accounted for the largest proportion of the variance, about 19.5% of the variance. The second and third factors

only account for about 1.8% and 1.0% respectively. Overall, this is a relatively low proportion of variance, meaning that even if the factors can be interpreted, there is still much of the variance unexplained.

A table with the factor loadings for the initial factor matrix is given in Appendix C. According to Stevens (2002), with a large sample size, as is the case here, even small factor loadings would be statistically significant. He states that testing at the alpha=0.01 level, for a sample size of 1,000, factor loadings of 0.16 would be considered statistically significant. This does not imply that these loading would be practically significant. He goes on to recommend for large sample sizes, factor loadings of 0.40 or more should be used.

Looking at the three factors, one can see that for the first factor, all but five variables had loadings of more than 0.40. The second factor only had two items with factor loadings greater than 0.40 and the third had no significant factor loadings. In order to see if a better model could be found, I conducted another factor analysis, this time using the Varimax rotation technique in order to get a better distribution of variable across factors. Rotation methods in factor analysis are used to improve the factor structure. Stevens (2002) states that the Varimax rotation method was designed to "clean up the factors" (p. 391). Basically this rotation method will force each factor to load high on a smaller number of variables and low on the others (Stevens, 2002).

This time the rotated factor pattern yielded three factors that explained approximately 8.4%, 7.1% and 6.8% of the variance. In the rotated factor analysis, the first factor yielded 25 items with factor loadings greater than 0.4; the second factor yielded sixteen items and the third factor sixteen items. Table 4.4 below shows the item numbers that had significant loadings for each factor.

	Item numbers
Factor 1	1, 7, 8, 9, 12, 16, 17, 18, 19, 20, 21, 22, 23, 25, 27, 29, 30, 32, 36, 39, 41, 43, 46, 48, 51
Factor 2	5, 14, 16, 29, 31, 34, 37, 40, 41, 42, 44, 54, 56, 57, 59, 60
Factor 3	3, 4, 7, 11, 15, 22, 23, 25, 28, 43, 45, 47, 54, 55, 58, 60

Table 4.4: Item numbers with factor loadings greater than 0.40

Interpretation of the above factors proved difficult. Initially I expected the factor loadings to represent various broad content areas such as algebra and geometry. Looking at the items that loaded on Factor 1, for instance, we can see that the items are not all of the "algebra" items or the "geometry" items. In fact, I was not able to find any relationship or pattern to the factor loadings for each of the three factors. There is nothing obvious in a surface level examination of what each of the items is asking. The items for each factor seem to be assessing very different skills. For instance items 7, 8, and 9 all load on factor 1, however item 7 asks students to find how much fence is needed to fence in a garden, item 8 asks students to find the x-intercept of a graph and item 9 asks students to find how many cubes would fit in a box.

Since I could not see a pattern, I examined the items through the interview and content area specialist data I collected. Again, I could find no real pattern. Since I could not find any patterns by examining the qualitative data, I decided to look at the quantitative data. I looked at the differential item functioning data, which will be discussed in detail in the next section, and the p-values for each item (p-values are an indicator of how difficult an item is). Yet again, I could not find a strong pattern to explain the factor loadings. There does not appear to be an obvious relationship among the items within each of the factors. This result is not entirely unexpected since the TAKS is designed to be a multidimensional test, assessing student performance on many different interrelated objectives in both algebra and geometry. Also, in both the initial factor matrix and the rotated factor matrix, only about 22% of the variance is accounted for from the factor analysis. This means there is considerable variance still left unexplained.

Differential Item Functioning (DIF) Analysis

Differential Item Functioning analysis is performed in order to see if students from various subgroups of the population perform differently on specific items from an

assessment. Each of the sixty items was analyzed for evidence of differential item functioning (DIF) across ethnicity. The data used for this analysis was the random sample of 4,340 students test responses provided by TEA. The DIF analysis was conducted using SPSS statistical software and the Mantel-Haenszel statistic. DIF analysis is conducted on matched groups. In this case, the total test score was used as a matching variable. Since the minimum score needed to pass on this test was 24 out of sixty, I decided to divide the data into three groups as follows: students with total scores from 0 to 23 were placed in Group 1, students with scores between 24 and 41 in Group 2 and students with scores between 42 and 60 in Group 3. Group 1, therefore, consists of all of the students who did not pass the exam. Within groups, students were compared across ethnicities using White students as the comparison group. No DIF analysis was conducted with Asian or Native American subgroups because the number of students in each group was too small.

The graphs below are given to illustrate exactly what DIF is. Both graphs show the proportion of students who answered the item correctly across score groups and ethnicities. The first graph illustrates an item with no significant DIF while the second graph illustrates an item with significant DIF.

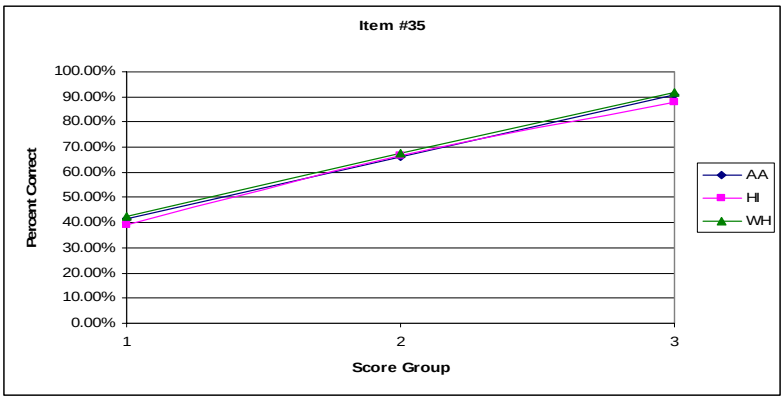


Figure 4.6: Example of no significant DIF

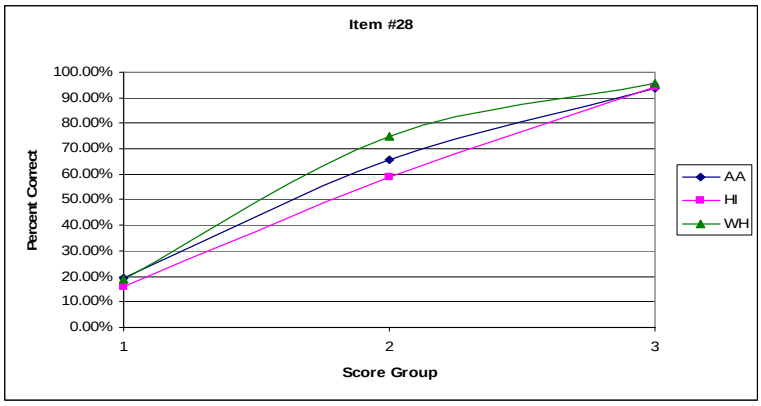


Figure 4.7: Example of significant DIF

In Figure 4.6, we see that there is virtually no difference between the performance of each of the subgroups for each of the three score groups. Figure 4.7, on the other hand, shows a distinct difference in performance across ethnicities for score group number 2. Although there is a noticeable difference between the percent of White students who answer correctly and African American students who answer correctly, and White students answering correctly and Hispanic students answering correctly, the difference between White students and Hispanic students is the most pronounced, and therefore has a greater statistical significance. An item can exhibit significant DIF across different score groups and ethnicities, however only the highest level of DIF is reported as "the DIF" for the item. Out of the sixty items, 39 exhibited a statistically significant level of DIF (at the .05 level). Of those 39, 14 items were significant at the alpha less than .000 level. (See figure 4.8 below) Since multiple comparisons are run in order to conduct a DIF analysis, the Type I error rate is significantly elevated, therefore, for this discussion, I will only focus on the 14 items with significance levels less than 0.000 level.

ITEM	DIF Significance Level	Advantaged Ethnicity/Disadvantaged Ethnicity	Score Group	TEA-Stated Objective
1	p<0.000	White/Hispanic	3	Functional Relationships
7	p<0.000	White/African American	1	Geometric Relationships and Spatial Reasoning
9	p<0.000	White/Hispanic	3	Mathematical Processes and Tools
11	p<0.000	White/African American	2	Properties and Attributes of Functions
15	p<0.000	White/African American	2	2-D and 3-D Representations
22	p<0.000	White/African American	2	Mathematical Processes and Tools
28	p<0.000	White/Hispanic	3	Mathematical Processes and Tools
30	p<0.000	White/African American	3	Mathematical Processes and Tools
32	p<0.000	White/Hispanic	3	Mathematical Processes and Tools
43	p<0.000	White/African American	2	Measurement
47	p<0.000	White/African American	2	Properties and Attributes of Functions
49	p<0.000	White/African American	3	Percents, Proportions, Probability and Statistics
52	p<0.000	White/African American	3	Measurement
58	p<0.000	White/African American	2	Mathematical Processes and Tools

Figure 4.8: DIF significance levels

In ten out of the fourteen cases, the disadvantaged ethnic group was the African American students. Hispanic students were the disadvantaged group in the remaining cases. Also, the DIF was most prevalent in score groups 2 and 3. Item 7 was the only instance where the most significant DIF occurred in score group 1. It is also interesting to note that in each of the four cases where the Hispanic students were the disadvantaged group, the most significant DIF occurred in score group 3, the highest scoring students. Three of the four items, items 9, 28 and 32, were classified by TEA under the Mathematical Processes and Tools objective. Specifically, items 9 and 28 are the two items under the sub-objective for developing and applying a problem solving strategy.

TEA states the sub objective as follows "The student is expected to…select or develop an appropriate problem solving strategy from a variety of different types, including drawing a picture, looking for a pattern, systematic guessing and checking, acting it out, making a table, working a simpler problem, or working backwards to solve a problem" (TEA, p. 6). Item 32 falls under the sub-objective for communicating mathematical ideas that states, "The student is expected to…communicate mathematical ideas using language, efficient tools, appropriate units and graphical, numerical, physical, or algebraic mathematical models," (TEA, 2004d, p. 6). This seems to indicate a pattern in the types of items exhibiting the highest degree of DIF for the Hispanic students. This pattern should be investigated further in order to ascertain if, in fact, it exists in other TAKS administrations and at lower grade levels.

For the ten items where the African American students were the disadvantaged group, I did not see any similar patterns or trends. Even so, there is something to be said about the fact that in ten out of the fourteen items where DIF was the most prevalent, it was the African American students who were the most significantly disadvantaged group. Again, further analysis should be conducted in order to determine if this is a pattern across grade levels and over various administrations of TAKS.

As stated at the beginning of this section, an item can exhibit significant DIF in multiple categories; however, only the category with the most significant DIF is reported. Of the fourteen items listed in Figure 4.8, eight showed significant DIF ($p<0.01$) in two or more categories. I say this in order to point out the fact that DIF analysis will not always show every instance where differences in performance exist.

Comments About Quantitative Results

The above analyses help to provide a more objective look of the items on the test. Both the factor analysis and the DIF analysis were exploratory in nature, conducted in hopes of finding areas of the test to investigate further. I believe the DIF analysis conducted here is one place to begin investigating differences in performance by various ethnic sub-groups. The DIF analysis helps to identify items that seem to be particularly

problematic to certain groups. Those items should be examined more closely in order to ascertain the reasons behind the differences in performance.

Chapter 5: Discussion

As stated earlier, Messick defines a multifaceted concept of validity where in order to assess the validity of a test, one looks at multiple lines of evidence. According to Messick, a validity study should take into account three main considerations: the appropriateness of the test, the meaningfulness of the test, and the usefulness of the test. Each one of these should be examined through empirically grounded construct interpretation. He states, "there is no way to judge responsibly the appropriateness, meaningfulness and usefulness of score inferences in the absence of evidence as to what the scores mean," (Messick, 1989, p. 35). This is illustrated in a diagram, taken from Messick (1989), reproduced in figure 5.1 below.

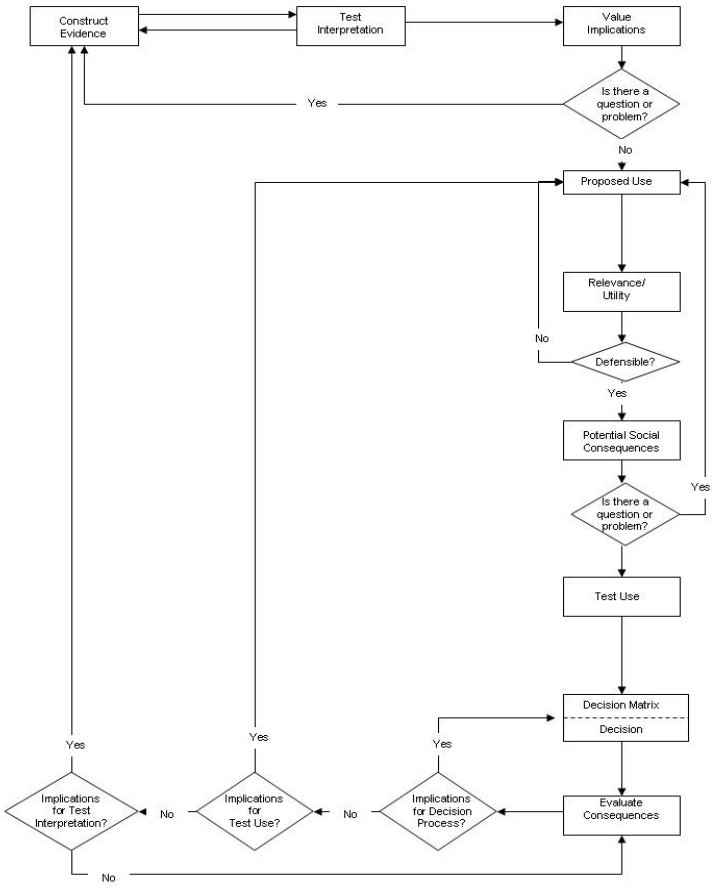


Figure 5.1: Validity model

Following the flow chart depicted in Figure 5.1, we can see that the relationship between test interpretation and construct validity is cyclic. If we cannot find consistent construct measures in an assessment or if we cannot verify that a particular assessment measures the constructs it was designed to measure, we cannot make valid interpretations about test scores. From here, we are caught in a feedback cycle until the constructs can be validated; therefore we cannot assess the social consequences of the test. If we cannot validate the underlying constructs of the test, other discussions of validity are moot. The purpose of this study is to examine the validity of the math portion of the 2004 TAKS, so we must begin with a careful examination of the underlying constructs of the test.

TEA lists a very well defined set of objectives the TAKS is meant to assess; therefore we know the intended constructs for the test, item by item. This study was designed to look at the various aspects of validity through multiple lines of evidence, namely statistical analysis of item responses, surveys of content area specialists and interviews with students on the items. Each aspect of this study adds to the understanding of the actual underlying constructs of the exam, versus the intended. Once we understand what these constructs are, we can better examine the appropriateness, meaningfulness and usefulness of the test

To summarize, in order to assess the validity of a test, we must make a comparison of what the intended outcomes are versus the actual outcomes of the test. The general intent of this test, as stated by TEA is to "better reflect good instructional practice and more accurately measure student learning" (TEA, 2004d, p. 1). A detailed description of the specific objective each item is intended to measure is given in the TAKS information booklet; therefore, we can use this information to better examine the intended versus the actual constructs of the items.

Discussion of Results

As stated earlier, each piece of this study was designed to give a certain perspective on the test. The content area specialist surveys were intended to provide some

insight into the underlying constructs the test appears to measure from an expert perspective while the student interviews were intended to provide some insight into what the items actually measure. The factor analysis was conducted in order to gain a more objective view of the constructs of the test. The DIF analysis allows us to see the differences in performance of various sub-groups on the items, thereby providing another perspective on the interpretation of the scores in relation to the underlying constructs.

In general, the results of the content area specialist surveys did not match the TEA-stated objectives. While some useful insight was gained into some of the items, further work needs to be done before specific conclusions about the underlying constructs of the test can be drawn. As stated earlier, the fact that the content area specialists were not able to verify the TEA-stated objectives does not in and of itself invalidate the constructs of the test. The discrepancies could very well be due to a difference in interpretation of the broad content area categories listed in the survey used. Nonetheless, the fact that there was such a low level of agreement between the objectives stated by the content area specialists and the TEA stated objectives could be indicative of some underlying problem with the way TEA defines the objectives for the test. More detailed surveys and/or interviews with a greater number of content area specialists would be needed in order to make any more definitive conclusions.

The factor analysis was inconclusive as a means of validating the internal structure and underlying constructs of the test in that the test did not factor across objectives. After an orthogonal rotation of the factors, one can see significant factor loadings for the items across three factors; however, the meaning of the factors is not readily apparent. This does not necessarily imply that the test does not measure the stated objectives. Although the test is designed to measure ten different overarching objectives, the sub-objectives for each of those are quite diverse. For example, 25 of the items are designed to assess Algebra knowledge, but within those items, 24 different sub-objectives are stated by TEA. The way this test was designed, the majority of sub-objectives are assessed with only one item. With a test that is designed to be multidimensional, such as the TAKS, it is not surprising that the factor analysis does not produce objective-level

factor loadings. The factor analysis does show that there are some possible correlations between groups of items, however the reason for the correlations is unclear. This could be an artifact of the way TEA defines the overarching objectives for the test or proof that the test does not in fact measure the intended objectives. More work would need to be done in order to make any definitive conclusions.

Through the DIF analysis we can see that there appear to be certain items on the test that are particularly problematic for specific groups. In fourteen of the sixty items, significant DIF was detected, mostly with respect to African American students as compared to White students, and mostly within the highest scoring groups. The items varied by intended objective so no generalizations could be made as to specific constructs where groups of students were disadvantaged, with one notable exception. In three of the four items where Hispanic students were the most significantly disadvantaged, each of the items was designed to measure students' uses of mathematical processes and tools; in each instance, it was the highest scoring Hispanic students who were significantly disadvantaged. I cannot conclude with certainty that the differences in performance on any of these items are due to flaws in the test construction, and not to some other factor such as true differences in content knowledge. Although at this time we cannot conclusively attribute these differences to a specific factor, the fact that there are significant differences is noteworthy and warrants further research. In particular, interviews should be conducted with Hispanic students to see what might be going wrong on these items. O'Neill and McPeek (1993) state that items that show high levels of DIF should be examined closely to see if the difference comes from factors in one of three broad areas: 1) surface level features, 2) true differences in the group's knowledge or 3) the criteria used for matching (O'Neill & McPeek, 1993).

While the previously mentioned components of the study each yielded some insight into the test, the true picture of the actual constructs measured by the test comes from the results of the student interviews. Even if the content area specialist surveys or the factor analysis were to verify the TEA-stated objectives, the student interview data would still supercede those results. The true test of what construct an item measures is

how the test takers attempt to solve the item. In numerous cases, the students solved the problems using test taking strategies rather than actual content knowledge. In problems 2, 26 and 27 the students were able to use clues from the problem itself to help them answer the question rather than working those problems out mathematically. In items 42 and 59, students were able to solve the problems by using the graphing calculator in order to get the correct answer without showing a true understanding of the mathematics behind their solutions. Out of all items included in the student interviews, there were only two in which the students attempted to solve the problem first and then looked for the correct answer among the choices listed; these were items 40 and 60. In the other problems, the general method was to eliminate choices until only one or two remained. Many students were stumped with problem 21 which required a fill-in-the-blank response. They were unable to get clues from the answer choices and instead had to try to solve it themselves. Few interviewees were able to solve this problem correctly, which corresponds to the student performance on this item from the entire sample. Item 21 is the most difficult item on the exam, with a p-value of 0.30. Further investigations should be conducted in order to see if this is the most difficult item because it requires advanced reasoning or simply because multiple-choice test-taking skills cannot be applied. One possible method for examining this type of item would be to give students both the open-ended form of the question and the multiple choice form in order to see if the difficulty lies in the content of the question or in the format. Kazemi (2002) conducted a study similar to this. She interviewed 90 fourth grade students on selected math problems. For each type of question, she created an open-ended problem and a corresponding multiple-choice version where she changed the numbers and the context but not the actual intended concept. Some students received the open-ended items first and then the multiple-choice while others received the multiple-choice items first. This type of methodology could be used to address the issue with the free response problem on the 2004 TAKS.

The issue of calculator use should not be discounted either. The fact that the students interviewed relied on the graphing calculator to answer several of the questions illustrates that constructs other than content objectives are being measured. Item 59 is a

good example of this point. All of the students who used the graphing calculator in order to graph the answer choices answered correctly. The one student who just went by what he "knew" answered incorrectly. This was also an item that was problematic for the content area specialists. The objective measured as well as the difficulty of the item seems to change depending on whether or not a graphing calculator is used. Cohen and Kim (1992) state, "if a particular item changes in difficulty or discrimination when a calculator is used, it is likely that the problem being posed to the examinee is different than when a calculator is not used," (p. 304). According to Meel (1997) "questions may exist on a test resulting in students' scores that are reflective of their understanding of the tool rather than of the material to be assessed" (p. 171). In the interviews I conducted, all of the students seemed proficient in the use of the graphing calculator; however, I am sure this is not the case in general. Some students may not have the same calculator skills and would therefore possibly be at a disadvantage on items such as these. Meel (1997) further states that "this technological edge results in an undue advantage for the student with the technological superiority" (p. 152). He goes on to state, "items should be examined to determine if the difficulty level or objective changes when a calculator is used," (p. 172). If this is the case, "the assessor should decide whether the change is appropriate for the goals of the assessment," (Meel, 1997, p. 172). More work should be done focusing on the ways in which students use calculators in high-stakes assessments such as TAKS.

Another issue that warrants further examination is the issue of language and vocabulary in mathematics tests. One instance was cited earlier where one of the students interviewed did not know what depreciated meant in item 47 and therefore could not answer the question. This student was not an English Language Learner, but the vocabulary was certainly a barrier for her. Another instance was with item 19. Two students stated that they did not know what "sequins" were. One student was able to work through and get the correct answer, but the other stated that since he did not know what sequins were, he could not solve the problem and just guessed. My sample size was small, both in students and number of items studied, and did not include any English

Language Learners however I still encountered instances of language barriers in two of the problems. In both instances, the troublesome words were not math terms, but rather just part of the set-up of the context of the word problem. If I found students who had trouble with these items in my small sample, it is reasonable to say that there are probably many other students who had similar troubles with these as well as other items. Abedi and Lord (2001) studied the language factor in math tests and found that "unfamiliar or infrequent vocabulary and passive voice constructions may affect comprehensions for certain groups of students and that average and low-achieving students may be at a relatively greater disadvantage in answering mathematics items with complex language" (p. 232). They found that if the items were modified to include simpler language, the students were better able to answer the items. As with the issue of calculator use, items with more complex language may be assessing more than just the content objective for which they were designed. Abedi and Lord (2001) state these cases "strongly [suggest] that factors other than mathematical skills contribute to success in solving word problems" (p. 220). These issues should be investigated for various sub-groups of students, not just English Language Learners.

Limitations of the Study and Areas for Future Research

One of the major limitations of this study is in the statistical analysis. The DIF analysis has a significantly elevated Type I error rate; however, it is the nature of DIF analysis to involve multiple tests and comparisons, therefore when conducting DIF on a sample, the elevated error rate is inevitable. I tried to mitigate the effects of this type I error rate by using a very conservative significance level. The obvious way to fix error in the analysis is to follow up with DIF analysis on the entire population of students taking the 2004 TAKS exit level test. However the purpose of this analysis was exploratory in nature, conducted in order to identify items that were potentially problematic, which has been accomplished. Further work needs to be done in order to determine the causes of the differences in performance.

Another limitation of this study is the data collected from the content area specialists. Although feedback was gathered from both high school and university level content area specialists, including mathematics education specialists as well as research mathematicians, the scope of such feedback was still very much limited. However, the data collected from these specialists could be used to help create improved follow-up surveys to be administered to other content area specialists. The follow-up surveys could include more detailed questions about the items where discrepancies existed between the content area specialists' opinions and the TEA-stated objectives in order to determine if the differences are true differences or if they are artifacts of the way the objective and sub-objectives are defined.

The final limitation of this study is that all of the student interviews were conducted with students from one school. The level of diversity in the interviewees was high as far as gender and ethnicity are concerned; however no English language learners were interviewed. The results would be strengthened by conducting interviews with students from various schools across the state, focusing on getting a representative sample of not only types of students but types of schools as well (rural, urban, suburban, low-performing, high-performing, etc.). Also, interviews should be conducted on each of the items rather than the subset of items used in this study. The interviews conducted in this study yielded good insight into how students solved the problems so it would be helpful to have that information on all of the items.

Final Remarks and Conclusions

In regards to the mathematics portion of the exit level TAKS, I cannot say that the assessment as a whole truly measures the constructs it was designed to measure. In some instances, the intended TEA-stated objectives were validated by the student interview results such as with problems 21 and 22; however, in many cases they were not. Even in instances where the content area specialists confirmed the TEA-stated objectives, the student interview data showed other objectives were truly being assessed. After conducting the one-on-one student interviews, I would hypothesize that the TAKS is

more a measure of calculator skills, problem solving ability and test taking skills rather than algebra and geometry content knowledge.

One of the most important things I learned from this study is that with multiple-choice items, we cannot tell how a student came to his or her conclusion about which answer choice to pick just by looking at the results. This is a major problem with multiple-choice tests in general, not just with the TAKS examination. Without extensive interview data from test-takers, no one can know what objective an item is truly testing, therefore making validation of the item and the test as a whole nearly impossible, at least by Messick's standard for validity.

Appendix A: CAS Protocol

For each item, please provide the following information:

1) Correct answer choice – if a correct answer choice is not given, make note of what a correct answer would be.

2) General mathematical objective measured by item. In this section, I want your response for the primary objective measured by the item. This list of objectives is intended to serve as guide. It is by no means intended to be overly restrictive. If you do not feel that the objective measured by the item can be expressed with one of the listed objectives, feel free to select choice "Z" and write in another.

(A) Functional Relationships
(B) Properties and Attributes of Functions
(C) Linear Functions
(D) Linear Equations and Inequalities
(E) Quadratic and Other Non-linear Functions
(F) Geometric Relationships and Spatial Reasoning
(G) 2-D and 3-D Representations
(H) Concepts and Uses of Measurement and Similarity
(I) Percents, Proportions, Probability and Statistics
(J) Mathematical Processes and Tools/Application of Mathematics
(Z) Objective not found in above list (please specify)

3) Cognitive level of item. Here are the categories you might use to classify each of the TAKS items. They are not mutually exclusive and you should check all that apply, but focus on what you would consider to be the most important or primary feature of the item.

(1) Requires recall of fact or vocabulary word. Example: Which of these figures is a parallelepiped?
(2) Requires direct application of formula given in the formula list at the beginning of the test or within the problem itself. Example: The base of a triangle is 6 cm and its height is 10 cm. What is the area of the triangle?
(3) Requires direct application of standard formula not given in the list.
(4) Requires manipulation of formula. Example: The area of a triangle is 60 cm^2. Its height is 10cm and the length of its non-base side is 5 cm. What is the length of its base. (Students would have to rearrange the formula as given to solve for the base and would have to ignore the extraneous information about the non-base side. This would be checked in addition to either 2 or 3).

(5) Requires application of conceptual knowledge. Example: In the drawing shown here, which two angles are equal? If there are 30 students in your class, what is the probability that another student will have the same birthday that you do?
(6) Requires development of an original procedure (problem solving). Example: The following chart shows Geiger counter readings vs. time. Predict what the reading would be at a time not shown.
(7) Requires synthesis of information from more than one source.

In the "Additional Comments" column make note of any item that stands out for some reason along with a brief explanatory statement.

Please record responses in the given Excel spreadsheet and then email the file to Erica Slate Young at eslate@mail.utexas.edu by April 8, 2005.

If you have any questions, please feel free to email me at the above address or call me at 254-368-9510.

Appendix B: CAS Survey Results

	Correct Answer				Primary Objective				Cognitive Level(s)			
Item	CAS 1	CAS 2	CAS 3	CAS 4	CAS 1	CAS 2	CAS 3	CAS 4	CAS 1	CAS 2	CAS 3	CAS 4
1	A	a	a		A	c	a		3	5	5	
2	H	h	h	H	I	i	i	I	5	5	3,5	5
3	D	d	d		A	g	g		6	5	1?	
4	J	j	j		G	e	g		6	6	3	
5	C	c	c	C	A	f	e	A	4	6	6	6
6	J	j	j	J	F	g	h	H	5	5	1	1,5
7	A	a	a		F	f	f		7	2	2	
8	J	j	j	J	B	j	c	C	4	5	5	4
9	B	d	b	B	F	f	g	F	7	3	2,4	5,6
10	J	h	f	J	F	f	f	F	5	5	5	1
11	D	d	d		A	c	a		5	5	3	
12	H	h	h	H	I	j	f	Z - trig	7	2	2,4	3
13	D	d	a	D	BD	c	c	C	1	5	1	2
14	F	h	f		I	i	i		5	5	5	
15	D	d	d		F	g	f		6	5	6?	
16	J	j	j		BD	b	d		4	7	5	
17	A	a	a		BD	c	c		4	7	5	
18	F	f	f		F	f	j		1	5	5	
19	B	b	c	B	F	f	h	F	2	5	2,3	2,4
20	H	h	g		D	d	d		7	5	2,4	
21	140	50	140	140	F	f	f	F	7	7	5,7	5,7
22	F	f	f	F	F	f	h	F	2	7	3	2,4
23	B	b	b		F	f	h		5	7	2,4	
24	F	f	h		F	g	g		1	5	1	
25	D	d	d		F	j	f		2	5	2	
26	F	f	f		F	f	f		5	7	2,4	
27	B	a	b	B	B	e	b	B	5	5	1	1
28	J	g	j	J	I	a	j	I	5	7	6	5,6
29	C	c	c		C	c	d		4	7	5	
30	F	f	f	F	H	f	c	F	5	5	2	5
31	D	d	d		A	e	a		4	5	6	
32	G	g	g	G	J	d	z	A	5	7	5	5
33	D		d	D	Z		z	J	5		1	5
34	H		h	H	B		e	E	7		1	1
35	B		b		I		i?		7		7?	
36	J	j	j		A	g	a		4	7	6	
37	A	a	c	A	E	e	j	E	4	4	3	2
38	J	j	j		F	g	f		5	5	6	
39	B		b		I		i		5		3	
40	H		h		D		d		5		5	
41	A		b	A	J		j	B	4		1	4
42	G		g		E		e		7		3,6?	
43	D	d	d		F	f	h		7	7	2	
44	F	f	f	F	B	c	d	D	4	7	2,4	2
45	D	c	d	D	A	c	a,i	C	6	5	3,4	2,7
46	F	f	f		F	f	h		7	2	2,4	
47	D	d	d		G	c	c		6	7	6	
48	H	h	h		B	d	d		4	7	2	
49	B	b	b	B	A	c	a	A	6	5	6	7
50	J	j	j		GF	f	f		7	5	1	
51	A	a	d	A	F	f	h	F	2	2	2,4	2,4
52	J	j	j	J	D	f	j	F	3	5	2,3,4	4,7
53	D	d	a		D	d	d		5	5	6	
54	J	j	j	J	J	f	h	exponents	1	4	2	1,2
55	B	b	b		A	c	a		5	5	3	
56	F	f	f		F	f	g		1	2	2	
57	D	d	d		BC	d	d		1	7	1	
58	H	h	h	H	I	j	z	I	1	7		5
59	D	b	a	D	BE	e	e	E	1	7	5	1
60	G	g	g		D	d	d		4	7	2	

Appendix C: Factor Analysis Results

Initial Eigenvalues

	Eigenvalue	Difference	Proportion	Cumulative
1	20.18436	17.81107	0.33640	0.33640
2	2.37328	0.77460	0.03960	0.37600
3	1.59868	0.42073	0.02660	0.40260
4	1.17795	0.02617	0.01960	0.42220
5	1.15178	0.10597	0.01920	0.44140
6	1.04581	0.03763	0.01740	0.45890
7	1.00818	0.05021	0.01680	0.47570
8	0.95797	0.01648	0.01600	0.49160
9	0.94149	0.00369	0.01570	0.50730
10	0.93781	0.01335	0.01560	0.52300
11	0.92445	0.03223	0.01540	0.53840
12	0.89222	0.03751	0.01490	0.55320
13	0.85471	0.01536	0.01420	0.56750
14	0.83935	0.02702	0.01400	0.58150
15	0.81233	0.01379	0.01350	0.59500
16	0.79854	0.00846	0.01330	0.60830
17	0.79008	0.01088	0.01320	0.62150
18	0.77920	0.01230	0.01300	0.63450
19	0.76690	0.01374	0.01280	0.64730
20	0.75316	0.00645	0.01260	0.65980
21	0.74671	0.00858	0.01240	0.67220
22	0.73813	0.03744	0.01230	0.68460
23	0.70069	0.00604	0.01170	0.69620
24	0.69465	0.00766	0.01160	0.70780
25	0.68699	0.01129	0.01140	0.71930
26	0.67570	0.00910	0.01130	0.73050
27	0.66661	0.01084	0.01110	0.74160
28	0.65577	0.00228	0.01090	0.75260
29	0.65349	0.01548	0.01090	0.76340
30	0.63800	0.00960	0.01060	0.77410
31	0.62841	0.01263	0.01050	0.78460
32	0.61578	0.01460	0.01030	0.79480
33	0.60118	0.01887	0.01000	0.80480
34	0.58231	0.01213	0.00970	0.81450
35	0.57018	0.02023	0.00950	0.82400
36	0.54995	0.00525	0.00920	0.83320
37	0.54469	0.00467	0.00910	0.84230
38	0.54002	0.01198	0.00900	0.85130
39	0.52804	0.00494	0.00880	0.86010
40	0.52310	0.01325	0.00870	0.86880
41	0.50984	0.02064	0.00850	0.87730
42	0.48920	0.00546	0.00820	0.88550
43	0.48375	0.00193	0.00810	0.89350
44	0.48181	0.01360	0.00800	0.90160
45	0.46821	0.01776	0.00780	0.90940
46	0.45045	0.00681	0.00750	0.91690
47	0.44364	0.00490	0.00740	0.92430
48	0.43874	0.01266	0.00730	0.93160
49	0.42607	0.00582	0.00710	0.93870
50	0.42026	0.01577	0.00700	0.94570
51	0.40448	0.01436	0.00670	0.95240
52	0.39012	0.01702	0.00650	0.95890
53	0.37311	0.01181	0.00620	0.96510
54	0.36129	0.02880	0.00600	0.97120
55	0.33249	0.00429	0.00550	0.97670
56	0.32820	0.03194	0.00550	0.98220
57	0.29626	0.00236	0.00490	0.98710
58	0.29389	0.01253	0.00490	0.99200
59	0.28137	0.08320	0.00470	0.99670
60	0.19817		0.00330	1.00000

Initial (Unrotated) Factor Loadings

	Factor1	Factor2	Factor3
MR1	0.45203	0.11906	0.1885
MR2	0.275	-0.0952	-0.04482
MR3	0.53959	0.09854	-0.19146
MR4	0.53776	0.14955	-0.17466
MR5	0.66347	-0.08367	-0.00538
MR6	0.46535	0.06694	0.03131
MR7	0.65269	0.14733	-0.09548
MR8	0.5282	-0.00483	0.20091
MR9	0.49546	0.25551	-0.03781
MR10	0.35713	0.00705	-0.0346
MR11	0.62993	0.02846	-0.1294
MR12	0.4502	0.20435	0.10061
MR13	0.56716	-0.15451	0.15497
MR14	0.33216	0.19049	0.12755
MR15	0.47882	0.0634	-0.21668
MR16	0.64638	-0.10535	0.12447
MR17	0.61999	0.01101	0.14849
MR18	0.57221	0.22116	0.06617
MR19	0.49378	0.19718	0.09927
MR20	0.48711	-0.00256	0.26383
MR21	0.71954	0.17676	0.04442
MR22	0.6645	0.1795	-0.09243
MR23	0.69807	0.10077	-0.08144
MR24	0.49318	0.16395	-0.03904
MR25	0.72741	-0.00185	-0.0177
MR26	0.43518	0.14661	0.04324
MR27	0.46002	0.09331	0.19899
MR28	0.72934	-0.00365	-0.23853
MR29	0.69187	-0.14134	0.1411
MR30	0.64725	0.08109	0.10789
MR31	0.60899	-0.28768	0.07313
MR32	0.61752	0.27902	0.05757
MR33	0.28852	0.09241	0.03036
MR34	0.46228	-0.15805	0.18669
MR35	0.52499	-0.00083	0.02585
MR36	0.65567	0.0578	0.09028
MR37	0.59094	-0.19557	-0.01446
MR38	0.56701	0.06253	-0.08453
MR39	0.55911	0.12863	0.00379
MR40	0.58115	-0.22959	-0.04434
MR41	0.64144	-0.12619	0.15389
MR42	0.52123	-0.22354	0.23302
MR43	0.72562	0.05543	-0.08024
MR44	0.61323	-0.35148	0.14626
MR45	0.65018	0.01677	-0.07922
MR46	0.5434	0.23638	0.08729
MR47	0.65252	0.0221	-0.17176
MR48	0.61652	0.08974	0.01595
MR49	0.41701	0.16183	-0.01605
MR50	0.52731	-0.14649	-0.10939
MR51	0.69952	0.15588	0.04893
MR52	0.54887	0.07231	-0.05245
MR53	0.54519	-0.09273	0.01602
MR54	0.69148	-0.18285	-0.14654
MR55	0.53007	-0.04977	-0.16429
MR56	0.58736	-0.10113	0.08921
MR57	0.3606	-0.45637	-0.0237
MR58	0.63185	-0.03518	-0.32322
MR59	0.42743	-0.46708	-0.00779
MR60	0.73552	-0.35976	-0.23624

Variance explained by each factor

Factor1	Factor2	Factor3
19.598186	1.772003	0.989209

Rotated Factor Loadings (Varimax Rotation)

	Factor1	Factor2	Factor3
MR1	0.44906	0.19701	0.1165
MR2	0.0916	0.21806	0.17537
MR3	0.28942	0.1755	0.47219
MR4	0.32667	0.13739	0.46525
MR5	0.35993	0.43235	0.3616
MR6	0.34364	0.21046	0.24421
MR7	0.43871	0.22064	0.46441
MR8	0.43132	0.34127	0.12991
MR9	0.43441	0.06014	0.34623
MR10	0.20696	0.18374	0.22847
MR11	0.33793	0.29571	0.46122
MR12	0.45043	0.10763	0.20023
MR13	0.34462	0.4724	0.16623
MR14	0.38361	0.05966	0.1103
MR15	0.21807	0.16438	0.4535
MR16	0.40587	0.4698	0.24281
MR17	0.46947	0.36736	0.22629
MR18	0.51742	0.1537	0.29894
MR19	0.47258	0.13715	0.22453
MR20	0.44063	0.33105	0.05601
MR21	0.57151	0.26575	0.39206
MR22	0.46622	0.20206	0.47343
MR23	0.4474	0.28624	0.47115
MR24	0.37948	0.13204	0.33196
MR25	0.44023	0.39924	0.41981
MR26	0.37728	0.1326	0.22983
MR27	0.44472	0.22446	0.10847
MR28	0.32296	0.35157	0.60078
MR29	0.42212	0.52757	0.24911
MR30	0.50523	0.31698	0.28534
MR31	0.25013	0.58372	0.23597
MR32	0.57433	0.13038	0.34007
MR33	0.24827	0.09215	0.15026
MR34	0.29446	0.42452	0.08132
MR35	0.33856	0.29651	0.27156
MR36	0.48764	0.33631	0.30083
MR37	0.24555	0.47997	0.31145
MR38	0.34248	0.24382	0.39473
MR39	0.42271	0.20651	0.32838
MR40	0.20396	0.49506	0.32516
MR41	0.40642	0.49047	0.21287
MR42	0.31781	0.52014	0.06638
MR43	0.43894	0.3381	0.4786
MR44	0.25481	0.65385	0.16891
MR45	0.37042	0.32767	0.42979
MR46	0.51958	0.13038	0.26797
MR47	0.32574	0.30366	0.5074
MR48	0.4423	0.27218	0.34453
MR49	0.34329	0.09688	0.2704
MR50	0.18399	0.38389	0.36092
MR51	0.54944	0.27247	0.37402
MR52	0.35395	0.23324	0.35994
MR53	0.29279	0.37914	0.27679
MR54	0.24496	0.49529	0.47722
MR55	0.21234	0.29536	0.42203
MR56	0.35298	0.42581	0.23931
MR57	-0.05265	0.55982	0.15066
MR58	0.19932	0.30377	0.61069
MR59	-0.00897	0.60894	0.17331
MR60	0.12241	0.64112	0.54791

Variance Explained by Each Factor

Factor1	Factor2	Factor3
8.3945103	7.1699625	6.7949253

Appendix D: Interview Response Data

MR 2

# of responses:	23
# per answer choice	
F	1
G	1
H*	20
J	1
NO ANSWER	0

FINAL ANSWER	REASON	INTERVIEW #
F	you have to multiply the bottoms together so you would get something over 9	1
G	1 junior out of 3 people	9
H	2 out of 3 are sophomores	3,6,7,8,12,25,26,31,32,36
H	ennumerates possibilities and counts	4,5,10,11,13,17,19,20,28,29
J	2/3 is too obvious. Guesses	2

2 In a high school auditorium, 1 junior and 2 sophomores are seated randomly together in a row. What is the probability that the 2 sophomores are seated next to each other?

F $\frac{1}{9}$

G $\frac{1}{3}$

H $\frac{2}{3}$

J $\frac{5}{6}$

MR 9

# of responses: 23

# per answer choice	
A	2
B*	13
C	2
D	6

FINAL ANSWER	REASON	INTERVIEW #
A	Adds 8,6 and 2 then divides by 2	36
A	Volume of box (96) divided by volume of cube (8) = 8	11
B	Drew box (3D) then counted number of cubes	17
B	Drew box (3D) then drew cubes inside and counted	13,26,29,31
B	Volume of box (96) divided by cube (2), but remembers from class that is not right	25
B	Volume of box (96) divided by volume of cube (8)	1,5,6,7,8,9,18
C	Volume of box (48) divided by cube (2)	12
C	Volume of box (96) divided by cube (2), but remembers from class that is not right	3
D	Guesses	23
D	Volume of box (96) divided by cube (2)	2,4,10,20,32

9 How many 2-inch cubes can be placed completely inside a box that is 8 inches long, 2 inches wide, and 6 inches tall?

A 8

B 12

C 24

D 48

MR 14

# of responses:	23
# per answer choice	
F*	8
G	1
H	5
J	7
NO ANSWER	2

FINAL ANSWER	REASON	INTERVIEW #
F	seems like a good probability number	31
F	sets up correct ratios and multiplies	1,2,5,11,12,17
F	sets up incorrect ratio (10/36) and chooses answer closest	13
G	sets up incorrect ratios (4/26 and 4/10) and tries to multiply. When that doesn't work, cross multiplies	14
H	sets up correct ratios and adds incorrectly (4/26+5/10=9/36)	4,10
H	sets up incorrect ratios (4/26 and 4/10) and reduces the 4/26 to 1/4	32
H	sets up ratio 9/36=1/4	7,24
J	guesses	21
J	sets up correct ratios and adds	3,28
J	sets up correct ratios and then guesses	8
J	sets up correct ratios then converts to %	22
J	sets up incorrect ratio (1/10) and guesses	33
J	sets up incorrect ratios and multiplies (4/26 and 5/9)	9
No Answer	sets up incorrect ratios (5/26 and 5/10) and tries to add	23

14 Jamal has a game with 2 groups of tiles. The first group of 26 tiles is labeled with all the letters of the alphabet. The second group of 10 tiles is numbered 0 through 9. If Jamal draws 1 letter tile and 1 number tile at random, what is the probability that he will draw a letter in his name and an odd number?

F $\frac{1}{13}$

G $\frac{5}{52}$

H $\frac{1}{4}$

J $\frac{7}{26}$

MR 18

# of responses:	24
# per answer choice	
F*	11
G	3
H	2
J	8

FINAL ANSWER	REASON	INTERVIEW #
F	Incorrect def'n of supp angles (angles are equal). F is still a true statement	19
F	remembered def'n of supp angles so knew F was correct	1,2,7,10,13,17,22, 23,24,33
G	Incorrect def'n of supp angles	11
G	faulty logic	20
G	remembered def'n of supp angles. faulty logic	32
H	remembered def'n of supp angles. Chose H because it is a true statement	3
H	Supp. angles add to 90 so H is a true statement	6
J	remembered def'n of supp angles. Chose J because it is a true statement	12
J	Did not know supp. Angles. Chose J because it is a true statement	4,5,8,9,14,21,31

18 Given: Two angles are supplementary. The measure of one angle is 20° more than the measure of the other angle.

Conclusion: The measures of the angles are 70° and 90°.

This conclusion —

F is contradicted by the first statement given

G is verified by the first statement given

H invalidates itself because a 90° angle cannot be supplementary to another

J verifies itself because 90° is 20° more than 70°

MR19

# of responses:	13
# per answer choice	
A	1
B*	8
C	1
D	2
NO ANSWER	1

FINAL ANSWER	REASON	INTERVIEW #
A	sets up correct diagram uses PT incorrectly (wrong "c" chosen)to solve for length	33
B	sets up correct diagram uses PT to solve for length	18,19,21,22,24,31,32
B	does not know what sequins are so can't solve problem. Guesses	26
C	mostly a guess	15
D	guesses	27
D	line should be half of 12. Chooses closest	36

19 Megan is using an equilateral triangle as part of a design on a sweatshirt. Each side of the triangle is 12 inches long. Megan is sewing a line of sequins from the midpoint of one side of this triangle to the opposite vertex. Approximately how long will the line of sequins be?

A 13.4 in.

B 10.4 in.

C 8.5 in.

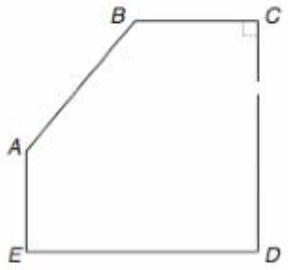
D 5.2 in.

MR 21

# of responses:	29
# per answer choice	
140*	9
130	9
OTHER	9

FINAL ANSWER	REASON	INTERVIEW #
140	partitioned picture into rectangle and quadrilateral	1,2,22
140	completed the "square" in the diagram	5,6,11
140	partitioned picture into triangles and a rectangle	17
140	uses "angles sum to 540 rule	24
140	partitions picture into triangles	32
130	assumed <A was the same as <B	12,13,14,23,25,28,31
130	complete the square in the diagram to show <A=<B	19
130	divides picturein half to show symmetry	26
180	guessed an angle bigger than 90	7
170	is not clear on which angle to solve for	36
145	completed the "square" in the diagram	4
128	guessed <A was close to <B	27
120	tried to partition picture into triangles	3
90	tried to partition picture into triangles	9
80	angles should sum to 480	29
65	divides 130 by 2	21
50	assumed <A and <B were supplementary	10

21 In the figure shown below, $\overline{BC}$ is parallel to $\overline{ED}$, and $\overline{AE}$ is perpendicular to $\overline{ED}$. The measure of $\angle ABC$ is 130°.



What is the measure of $\angle BAE$ in degrees?

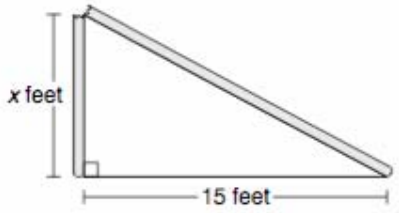
Record your answer and fill in the bubbles on your answer document. Be sure to use the correct place value.

MR 22

# of responses:	22
# per answer choice	
F*	13
G	7
H	1
J	1
NO ANSWER	0

FINAL ANSWER	REASON	INTERVIEW #
F	measured/estimated the side length	2,8,13
F	Plugs answer choices into PT to find which combinations add to 25	1,5,6,7,9,11,17,29
F	Reasons out that x must be less than 10	19,31
G	25-15=10	10,12,18,36
G	measured/estimated the side length	4,20,32
H	Uses PT, but finds wrong side	25
J	Set up PT inappropriately	3

22 A wooden pole was broken during a windstorm. Before it broke, the total height of the pole above the ground was 25 feet. After it broke, the top of the pole touched the ground 15 feet from the base.



How tall was the part of the pole that was left standing?

F 8 ft

G 10 ft

H 17 ft

J 20 ft

MR 26

# of responses: 24

per answer choice

F*	16
G	5
H	2
J	1

FINAL ANSWER	REASON	INTERVIEW #
F	2 is the only choice that is a factor of four	3,16,28
F	plugged number into SA formula then added 4	9,10,13
F	plugged number into SA formula then multiplied by 4	2,8,17,24,31
F	saw relationship between side and surface area	1,4,12,21,33
G	if one measurement is increased by 4 they all would	5,32
G	plugged number into SA formula then added 4. Could not solve and guessed G	7
G	plugged number into volume formula then multiplied by 4	22
G	plugged number into volume formula then raised answer to the fourth power	11
H	plugged number into SA formula then added 4. Could not solve and guessed H	6
H	Rules out extreme values and then rules out G b/c "they" would not repeat the same number	19
J	by guessing	23

26 If the surface area of a cube is increased by a factor of 4, what is the change in the length of the sides of the cube?

F The length is 2 times the original length.

G The length is 4 times the original length.

H The length is 6 times the original length.

J The length is 8 times the original length.

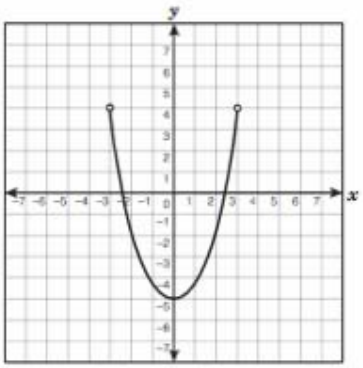
MR27

# of responses: 15

# per answer choice	
A	2
B*	13
C	0
D	0

FINAL ANSWER	REASON	INTERVIEW #
A	knows domain is x's cooses A b/c line is solid not dashed	4,10
B	knows domain is x's and open circles mean <	1,2,8,9,11,20
B	knows open circles mean <, does not state knowledge of domain	3,5,17,18
B	eliminated all others because they had a less then or equals to	7,12
B	is not sure about domain (guesses it is x) does recognize open circles	28

27 What is the domain of the function shown on the graph?



A $-3 \leq x \leq 3$

B $-3 < x < 3$

C $-5 < x \leq 4$

D $-5 \leq x < 4$

MR 32

# of responses:	20
# per answer choice	
F	1
G*	11
H	1
J	7
NO ANSWER	0

FINAL ANSWER	REASON	INTERVIEW #
F	solves for n, but chooses F	25
G	guessed	10
G	n-1 is less and n+1 is more so n is the middle	1,2,5,8,9,13,31,35
G	picked example numbers (20,21,22) and saw n was middle	17
G	solved equation for n	7
H	guessed	4
J	eliminated other choices and was left with J	12
J	sounds like the best answer	3,15,29,32
J	you would have to do something to the smallest and biggest to get the answer	27,36

32 Chase wanted to find 3 consecutive whole numbers that add up to 81. He wrote the equation $(n-1)+n+(n+1)=81$. What does the variable n represent in the equation?

F The least of the 3 whole numbers

G The middle of the 3 whole numbers

H The greatest of the 3 whole numbers

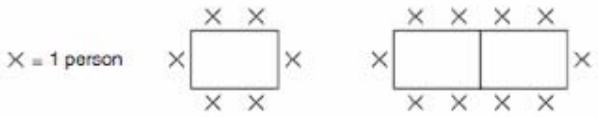
J The difference between the least and greatest of the 3 whole numbers

MR 36

# of responses:	14
# per answer choice	
F	1
G	4
H	0
J*	9
NO ANSWER	0

FINAL ANSWER	REASON	INTERVIEW #
F	six people per table	27
G	there are four seats with the ones on the end	15
G	it just makes sense	23
G	substitues number in but not correctly	32
G	4y for the number of people and 1 for the table	36
J	substitutes numbers into each answer choice	19,21,22,26,28,29,31
J	there are four seats per table and two extra on the ends (4x+2)	25,35

36 For a sports banquet Coach Mackey must use the rectangular tables in the school cafeteria. The diagram below shows the seating arrangements that Coach Mackey can use at 1 and 2 tables.



Which expression can be used to determine the number of people who can sit as a group if y tables are joined to form 1 long table?

F $6y$

G $4(y + 1)$

H $3(y + 1)$

J $2(2y + 1)$

MR 40

# of responses:	15
# per answer choice	
F	0
G	0
H*	15
J	0
NO ANSWER	0

FINAL ANSWER	REASON	INTERVIEW #
H	creates own equations and finds answer to match	1,2,3,4,5,7,8,9, 10,11,12
H	eliminates F and G chooses H b/c "more than" means to add	14,16,17,20

40 At a college bookstore, Carla purchased a math textbook and a novel that cost a total of \$54, not including tax. If the price of the math textbook, m, is \$8 more than 3 times the price of the novel, n, which system of linear equations could be used to determine the price of each book?

F $m + n = 8$
$m = 3n + 54$

G $m + n = 8$
$m = 3n - 54$

H $m + n = 54$
$m = 3n + 8$

J $m + n = 54$
$m = 3n - 8$

MR 42

# of responses:	13
# per answer choice	
F	0
G*	13
H	0
J	0
NO ANSWER	0

FINAL ANSWER	REASON	INTERVIEW #
G	Plugs into calculator, graphs the function looks at the graph to find the roots	13,21,31,33
G	Plugs into calculator, graphs the function then goes to table of values to find the roots	22,24,25,26,29
G	Plugs into calculator, graphs the function then uses TRACE function to find roots	28
G	Substitutes x-coordinate of answer choices into function see which gives a zero	23
G	Guessed. G just looked good (one of two whole number choices)	32,36

42 Which ordered pair represents one of the roots of the function $f(x) = 2x^2 + 3x - 20$?

F $(-\frac{5}{2}, 0)$

G $(-4, 0)$

H $(-5, 0)$

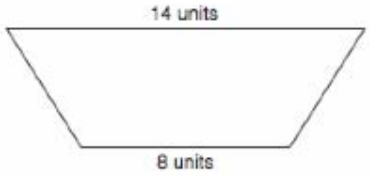
J $(-20, 0)$

MR 46

# of responses:	22
# per answer choice	
F*	14
G	1
H	2
J	5
NO ANSWER	0

FINAL ANSWER	REASON	INTERVIEW #
F	finds missing side (5) estimates height then uses formula	23,25
F	finds missing side (5) finds height (4) then uses area formula from sheet	2,3,5,7,9,10,26,31,35
F	finds missing side (5) finds height then calculates area of large rectangle and subtracts area of	11
F	plugs given info into area formula and then estimates height	29
F	Sets up the equation 32=8x, so h=4. Plugs into given area formula	12
G	Finds height (10) then plugs into equation	32
H	finds missing side (3) then works backwards through answer choices	19
H	plugs given info into area formula and gets A=16h can't find answer choice to and estimates	20
J	adds up all the sides (14+8+32) and 55 is the closest	36
J	estimates height (5) then uses given formula	13
J	finds missing side (5) then uses it as the height in area formula	4,8,28

46 The lengths of the bases of an isosceles trapezoid are shown below.



If the perimeter of this trapezoid is 32 units, what is its area?

F 44 square units

G 110 square units

H 88 square units

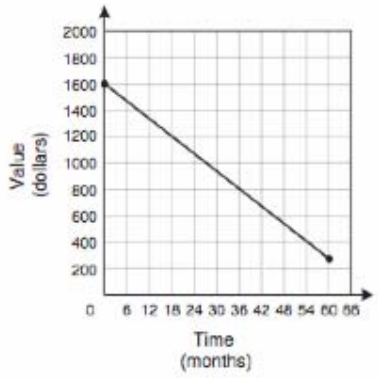
J 55 square units

MR47

# of responses:	13
# per answer choice	
A	1
B	2
C	1
D*	9
NO ANSWER	0

FINAL ANSWER	REASON	INTERVIEW #
A	Eliminates C and D b/c does not know what "depreciate" means. Eliminates B b/c the numbers are too high. Guesses A	15
B	Goes through answer choices to find one that works. Chooses B b/c 400 is half of 800.	27,29
C	Goes through answer choices to find one that works. Decides C is true.	32
D	Goes through answer choices to find one that works. Decides D is true.	13,18,19,20,22,23,28,31,35

47 The graph below shows the decrease in the value of a personal computer over a period of 60 months.



Which is a reasonable conclusion about the value of this personal computer during the time shown on the graph?

A Its value at 18 months was twice its value at 36 months.

B Its value at 36 months was half its value at 54 months.

C It depreciated $200 *every* 12 months.

D It depreciated $400 every 18 months.

MR 49
# of responses: 34
per answer choice

A	9
B*	14
C	11
D	0

FINAL ANSWER	REASON	INTERVIEW #
A	eliminates C and D, decides B does not work so A seems like the most rational one	6
A	found avg price per pound ($0.33) so A is true	8
A	eliminates C and D, decides B does not work somehow verifies A	11
A	eliminates B,C and D, misreads A (less than not more than) so chooses A	13
A	eliminates B, C and D so A must be the answer	21
A	Chooses A because $4 is in between $1.95 and $6.95	25
A	Guesses A	27,32
A	eliminates A, B, C and D chooses A because it is the closest	35
B	eliminates A,C and D so B must be the answer	2,3,19,22
B	eliminates C and D. B seems right because the price goes down the more you buy	4,31
B	eliminates C and D, chooses B for no real reason	7
B	eliminates A, B, C and D chooses B because it is the closest	10
B	eliminates A, C and D, chooses B for invalid reason	16,18
B	eliminates C and B just seems right	24
B	eliminates A,C and D chooses B because it could be true	28, 36
C	Assumes constant price per pound ($0.39) so C is true	5,9,12,14,20,26,29,33
C	eliminates A and B (they don't sound right). Guesses C	15
C	Assumes constant price per pound ($0.36) so C is true	17
C	eliminates A, B, C and D guesses C is answer	23

40 The table below shows the cost of fertilizer, depending on the amount purchased.

Cost of Fertilizer

Number of Pounds	Cost
5	$1.95
20	$6.95
50	$15.95
100	$28.95

Which conclusion can be made based on information in the table?

A The cost of 10 pounds of fertilizer would be more than $4.00.

B The cost of 200 pounds of fertilizer would be less than $57.00.

C The cost of fertilizer is always more than $0.35 per pound.

D The cost of fertilizer is always less than $0.30 per pound.

MR 51

# of responses: 23

per answer choice

A*	14
B	4
C	3
D	2

note: all students but one drew and labeled the correct figure

FINAL ANSWER	REASON	INTERVIEW #
A	draws correct picture and then estimates sides	13,23
A	uses PT then adds the sides	1,2,3,5,8,10,17,21,22 ,24,32,33
B	draws correct picture and then estimates sides	20
B	multiplies 30 and 40 then guesses	12
B	uses PT but does area instead	9
B	uses PT incorrectly. Guesses	15
C	draws correct picture but adds 30+30+40+40	4
C	draws incorrect picture and adds 30+30+40+40	35
C	uses PT incorrectly then adds 30+30+40+40	28
D	assumes 30-40-50 triangle	7,31

51 About how many feet of fencing are needed to enclose a rectangular garden with a 30-foot-long side and a 40-foot-long diagonal?

A 113 ft

B 133 ft

C 140 ft

D 160 ft

MR59

# of responses:	13
# per answer choice	
A	1
B	0
C	0
D*	12
NO ANSWER	0

FINAL ANSWER	REASON	INTERVIEW #
A	Remembered rule - larger numbers make parabolas wider. Did not check on calculator	8
D	Remembered rule - fractions make parabolas wider. Used calculator to verify.	11
D	Plugged anser choices into calculator, graphed each and found the widest	2,4,5,7,9,10,12,14,16,20,28

59 Which equation will produce the widest parabola when graphed?

A $y = 2x^2$

B $y = -6x^2$

C $y = -0.6x^2$

D $y = 0.2x^2$

MR 60

# of responses:	15
# per answer choice	
F	0
G*	15
H	0
J	0
NO ANSWER	0

FINAL ANSWER	REASON	INTERVIEW #
G	substituted 8625 into equation and solved for n	1,2,3,4,5,6,7,8,9,10,11,12,14,17,28

60 Ms. Barton determined that the total cost of her wedding, c, could be represented by the equation $c = 75n + 1500$, where n is the number of people attending the wedding. If Ms. Barton's wedding cost $8625, how many people attended the wedding?

F 135

G 95

H 115

J 75

Bibliography

Albrecht, S. F., & Joles, C. (2003). Accountability and Access to Opportunity: Mutually Exclusive Tenets Under a High-Stakes Testing Mandate. Preventing School Failure, 42(2), 86-91.

Alexander, V. L. (2003). High Stakes Testing: Its Intended and Unintended Consequences on Minority and Economically Disadvantaged High School Students. Unpublished dissertation, University of Texas, Austin, Texas.

Allen, M. J., & Yen, W. M. (1979). Introduction to Measurement Theory (1 ed.). Long Grove, Illinois: Waveland Press. Inc.

American Educational Research Association. (2000). AERA Position Statement - High-Stakes Testing in PreK-12 Education. American Educational Research Association. Available: http://www.aera.net/policyandprograms/?id=378 [2005, 1/20].

Amrein, A. L., & Berliner, D. C. (2002). An Analysis of Some Unintended and Negative Consequences of High-Stakes Testing. Arizona State University Education Policy Studies Laboratory. Available: www.asu.edu/educ/epsl/EPRU/documents/EPSL-0211-126-EPRU.pdf [2005, 1/15].

Amrein, A. L., & Berliner, D. C. (2003). The Testing Divide: New Research on the Intended and Unintended Impact of High-Stakes Testing. Peer Review, 31-32.

Darlington, R. (1997). Factor Analysis. Cornell Department of Psychology. Available: http://www.psych.cornell.edu/Darlington/factor.htm [2005, 1/20].

Frary, R. B. (2000). Higher Validity in the face of Lower Reliability: Another Look. Applied Measurement in Education, 13(3), 249-253.

Fuller, E. J., & Johnson, J. F. (2001). Can State Accountability Systems Drive Improvements in School Performance or Children From Low-Income Homes? Education and Urban Society, 33(3), 260-283.

Glaser, R., & Silver, E. (1994). Assessment, Testing, and Instruction: Retrospect and Prospect. Review of Research in Education, 20, 393-419.

Goertz, M., & Duffy, M. (2003). Mapping the Landscape of High-Stakes Testing and Accountability Programs. Theory into Practice, 42(1), 4-11.

Haney, W. (2000). The Texas Miracle in Education. Educational Policy Analysis Archives. Available: http://epaa.asu.edu/epaa/v8n41/ [2005, 1/20].

Heubert, J. P., & Hauser, R. M. (Eds.). (1999). High Stakes: Testing for tracking promotion and graduation. Washington, D.C.: National Academy Press.

Kohn, A. (2000). Burnt at the High Stakes. Journal of Teacher Education, 51(4), 315-327.

Kohn, A. (2001). Fighting the Tests. Phi Delta Kappan, 348-357.

Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. Theory into Practice, 41(4), 212-218.

McNeil, L. M. (2000). Creating New Inequalities. Phi Delta Kappan, 81(10), 729-734.

Measurement Research Associates. (2003). What Item Analysis Can Tell us About Item Quality.
Available: http://www.measurementresearch.com/media/itemanalysis.pdf [2005, 1/20].

Mehrens, W. A. (2000). Defending a State Graduation Test: GI Forum v. Texas Education Agency Measurement Perspectives From an External Evaluator. Applied Measurement in Education, 13(4), 387-401.

Messick, S. (1989). Validity. In R. L. Lin (Ed.), Educational Measurement. New York, NY: American Council on Education, Macmillan Publishing.

Nathan, L. (2002). The Human Face of the High Stakes Testing Story. Phi Delta Kappan, 83(8), 595-600.

O'Connell, P. J., McGuire, K., Middleton, R., Ruiz, R., Bellamy, G. T., & Bornfield, G. (2000). Agora: The Impact of High Stakes Testing. Journal of Teacher Education, 51(4), 289-292.

Phillips, S. E. (2000). G. I. Forum v. Texas Education Agency Psychometric Evidence. Applied Measurement in Education, 13(4), 343-385.

Plake, B. S. (2002). Evaluating the Technical Quality of Educational Tests Used for High Stakes Decisions. Measurement and Evaluation in Counseling and Development, 35(@), 144-152.

Ravitch, D. (2002). A Brief History of Testing and Accountability. Hoover Digest, 4.

Scheurich, J. J., Skrla, L., & Johnson, J. F. (2000). Thinking Carefully About Equity and Accountability. Phi Delta Kappan, 292-299.

Sclafani, S. (2001). Using an Aligned System to Make Real Progress for Texas Students. Education and Urban Society, 33(3), 305-312.

Shavelston, R. J., Roeser, R. W., Kupermintz, H., Lau, S., Ayala, C., Haydel, A., Schultz, S., Gallager, L., & Quihuis, G. (2002). Richard E. Snow's Remaking of the Concept of Aptitude and Multidimensional Test Validity: Introduction to the Special issue. Educational Assessment, 8(2), 77-99.

Shepard, L. A. (1993). Evaluating Test Validity. Review of Research in Education, 19, 405-450.

Sireci, S. G. (1998). Gathering and Analyzing Content Validity Data. Educational Assessment, 5(4), 299-321.

Sloane, F. C., & Kelly, A. E. (2003). Issues in High Stakes Testing Programs. Theory into Practice, 42(1), 12-17.

Smisko, A., Twing, J. S., & Denny, P. (2000). The Texas Model for Content and Curricular Validity. Applied Measurement in Education, 13(4), 333-342.

Strauss, A. & Corbin, J. (1990). *Basics of Qualitative Research: Grounded Theory Procedures and Techniques.* Newbury Park, California; London; New Delhi: Sage.

Texas Education Administration. (2004a). Technical Digest 2003-2004. Student Assessment Division.
Available: http://www.tea.state.tx.us/student.assessment/resources/techdig04/index.html [2005, 1/20].

Texas Education Administration. (2004b). Spring 2004 TAKS. Student Assessment Division.
Available: http://www.tea.state.tx.us/student.assessment/resources/techdig04/index.html [2004, 9/15].

Texas Education Administration. (2004c, August). TAKS Information Booklet Mathematics Exit Level - Revised. Student Assessment Division.
Available: http://www.tea.state.tx.us/student.assessment/taks/booklets/index.html [2004, 11/15].

Texas Education Administration. (2004d). Texas Assessment of Knowledge and Skills Answer Key. Student Assessment Division.
Available:http://www.tea.state.tx.us/student.assessment/resources/release/taks/2004/gr11takskey.pdf [2005, 9/25].

Texas Education Administration. (2004e). Texas Assessment of Knowledge and Skills Summary Report. Student Assessment Division. Available:http://www.tea.state.tx.us/student.assessment/reporting/results/summary/sum04/taks/g11.pdf [2005, 9/25].

Trueba, H. T. (2001). Conclusion: Polar Positions on the Texas Assessment of Academic Skills (TAAS). Pragmatism and the Politics of Neglect. Education and Urban Society, 33(3), 333-344.

Valencia, R. R., Valenzuela, A., Sloan, K., & Foley, D. E. (2001). Let's Treat the Cause, Not the Symptoms: Equity And Accountability in Texas Revisited. Phi Delta Kappan, 83(4), 318-321,326.

Varma, S. Computing and Using point Biserials for Item Analysis. Educational Data Systems. Available: www.eddata.com/resources/publications/EDS_Point_Biserial.pdf [2005, 1/20].

Vinovskis, M. A. (1998). Overseeing the Nation's Report Card. University of Michigan School of Public Policy. Available: http://nces.ed.gov/nationsreportcard/about [2005, 1/25].

Ward, C. A. (2000). G. I. Forum v. Texas Education Agency Implications for State Assessment Programs. Applied Measurement in Education, 13(4), 419-426.

Webb, N. L., Clune, W. H., Bolt, D., Gamoran, A., Meyer, R. H., Osthoff, E., & Thorn, C. (2002). Models for Analysis of the Impact of Systemic Initiative Programs--The Impact of Urban Systemic Initiatives on Student Achievement in Texas, 1994-2000. Wisconsin Center for Education Research, University of Wisconsin-Madison. Available: http://facstaff.wcer.wisc.edu/normw/technical_reports%20Page%201.htm [2004, 11/30].

Zenisky, A. L., Hambleton, R., & Robin, F. (2003). Detection of Differential Item Functioning in Large Scale State Assessments: A study Evaluating a Two-Stage Approach. Educational and Psychological Measurement, 63(1), 51-64.

Vita

Erica Rae Slate was born in Boone, North Carolina on April 11, 1973, the daughter of Lynda Slate Stanbery and the late James Raymond Slate. After graduating from Watauga High School in 1991, she attended Appalachian State University (ASU) where she earned a Bachelor of Science in Mathematics, Secondary Education in 1996 with a minor in music and an endorsement in Physics, and a Master of Arts in Mathematics Education in 1999. While earning her degrees, she taught Pre-Calculus at ASU and Algebra II and Calculus at Heavenly Mountain Ideal Girls' School. After completion of her Master of Arts Degree, she continued to teach at ASU and also taught various math courses at and Wilkes Community College.

In 2001, she enrolled at the University of Texas at Austin to begin work on a doctoral degree in Mathematics & Science Education. Her education at UT Austin has included experience as a graduate research assistant in several research areas including mathematics curriculum evaluation, systemic reform in mathematics education, after-school mathematics programs, and design experiments in the mathematics classroom. She was also a teaching assistant in the secondary teacher preparation program, UTeach. She is currently working at the United States Military Academy at West Point as an Assistant Professor in Mathematics.

Permanent address: 285 Cedar Rock Drive

Boone, North Carolina 28607

This dissertation was typed by the author.