

2011

Finding Relationships Between Multiple-Choice Math Tests and Their Stem-Equivalent Constructed Responses

Nayla Aad Chaoui
Claremont Graduate University

Recommended Citation

Chaoui, Nayla Aad, "Finding Relationships Between Multiple-Choice Math Tests and Their Stem-Equivalent Constructed Responses" (2011). *CGU Theses & Dissertations*. Paper 21.
http://scholarship.claremont.edu/cgu_etd/21

DOI: 10.5642/cguetd/21

This Open Access Dissertation is brought to you for free and open access by the CGU Student Scholarship at Scholarship @ Claremont. It has been accepted for inclusion in CGU Theses & Dissertations by an authorized administrator of Scholarship @ Claremont. For more information, please contact scholarship@cuc.claremont.edu.

**Finding Relationships Between Multiple-Choice Math Tests
And Their Stem-Equivalent Constructed Responses**

By

Nayla Aad Chaoui

A dissertation submitted to the Faculty of Claremont Graduate
University in partial fulfillment of the requirements for the
degree of Doctor of Philosophy in the Graduate Faculty of
Education

Claremont Graduate University

2011

APPROVAL OF THE DISSERTATION COMMITTEE

We, the undersigned, certify that we have read, reviewed, and critiqued the dissertation of Nayla Aad Chaoui and do hereby approve it as adequate in scope and quality for meriting the degree of Doctor of Philosophy.

Dr. Mary Poplin
Chair
School of Educational Studies

Dr. June Hilton
Committee Member
School of Educational Studies

Dr. Phil Dreyer
Committee Member
School of Educational Studies

Abstract of the Dissertation

Finding Relationships Between Multiple-Choice Math Tests
And Their Stem-Equivalent Constructed Responses

By

Nayla Aad Chaoui

Claremont Graduate University, 2011

The study takes a close look at relationships between scores on a Mathematics standardized test in two different testing formats - Multiple-Choice (MC) and Constructed Response (CR). Many studies have been dedicated to finding correlations between item format characteristics with regards to race and gender. Few studies, however, have attempted to explore differences in the performance of English Learners in a low performing, predominantly Latino high school. The study also determined relationships between math scores and gender and math scores and language proficiency, as well as relationships between CAHSEE and CST scores.

Statistical analyses were performed using correlations, descriptive statistics, and t-tests. Empirical data were also disaggregated and analyzed by gender, and language proficiency. Results revealed significant positive correlations between MC and CR

formats. T-tests displayed statistically significant differences between the means of the formats, with boys and English Only students having better scores than their counterparts. Frequency tables examining proficiency levels of students by gender and language proficiency revealed differences between MC and CR tests, with boys and English Only students earning better levels of proficiency. Significant positive correlations were shown between CST scores and multiple-choice items, but none were found for CST scores and constructed response items.

DEDICATION

To my husband Nabil, whose patience and many sacrifices are what allowed me to complete this work uninterrupted.

To my wonderful children, who thrived and finished their schooling while their mother was busy attending classes and focusing on her study.

To my students, who are my biggest fans, and who accompanied me on this journey. May this work inspire you to always believe in yourselves and reach for the stars.

And finally, to my parents, who instilled in me the love of learning, and taught me to appreciate culture and diversity.

ACKNOWLEDGEMENTS

I am indebted to my chairperson, Dr. Mary Poplin, whom I not only consider to be an invaluable source of information and guidance, but a very dear friend. I am eternally grateful for my committee members, Dr. June Hilton and Dr. Phil Dreyer, who graciously stepped in when I needed them the most. I would like to especially acknowledge Dr. Hilton, for her efforts in teaching me how to use the software to analyze the data.

I want to thank Mr. Stacey Wilkins, my principal, without whom this study would not have been possible; Mrs. Valerie Cordova, our secretary, who patiently accessed scores and provided me with the necessary data; and to Mrs. Glenda Vazquez Hermosillo, our assistant principal in charge of testing, who supplied me with vital information as well.

Finally, I would like to acknowledge my friends and colleagues, who were rooting for me and supporting me throughout this journey. Your kindness and consideration did not go unnoticed.

Table Of Contents

Dedication	
Acknowledgements.....	vi
Table of Contents.....	vii
List of Tables	ix
CHAPTER I INTRODUCTION	1
Overview.....	1
Theoretical Framework.....	3
Background.....	4
Significance Of the Topic.....	9
Research Questions.....	11
Methodology.....	14
Summary.....	17
CHAPTER II LITERATURE REVIEW.....	18
Overview.....	18
Constructed response Tests.....	18
Multiple Choice Tests.....	23
MC versus CR items.....	27
Gender.....	36
Ethnicity and Language.....	40
Scoring Rubrics.....	43
Validity.....	45
Guidelines for Writing MC Questions.....	47
Guidelines for Writing CR Questions.....	49
Effective Math Instruction.....	51
CHAPTER III METHODOLOGY.....	76
Overview.....	76
Data Set.....	77
Key Variables.....	80
Instrumentation.....	80
The CAHSEE Math Standards.....	87
The Math Test.....	90
The CMC Scoring Rubric.....	100
Procedures.....	102
CHAPTER IV RESULTS.....	104
Correlations between Percents of Correct Answers.....	104
T-tests for Gender.....	105
T-tests for Language.....	106
Proficiency Levels on MC and CR Tests.....	107-110
Pearson Correlations between MC and CR Scores.....	111
Pearson Correlations by Gender and Language.....	112

Pearson Correlations by Strand.....	113
T-tests for Gender.....	114
T-tests for Language.....	115
Pearson Correlations between CST and CAHSEE Scores.....	116
Pearson Correlations for Gender and Language.....	117
CHAPTER V CONCLUSION.....	119
Research Findings.....	120
Limitations of the Study.....	125
Implications of the Study.....	126
Appendix A.....	133
Bibliography.....	136

List of Tables

Table 1: Demographics of Students.....	78
Table 2: Content Standards of the Mock CAHSEE.....	87
Table 3: CMC Scoring Rubric.....	100
Table 4: Correlations between Percents of Correct Answers on MC and CR Items.....	104
Table 5: T-test Results for the Differences between Percents of Correct Answers (Gender).....	105
Table 6: T-test Results for the Differences between Percents of Correct Answers (Language).....	106
Table 7: Proficiency Levels of 9 th Graders on MC Test....	107
Table 8: Proficiency Levels of 9 th Graders on CR Test....	108
Table 9: Comparison of Proficiency Levels of 9 th graders on CAHSEE.....	108
Table 10: Proficiency Levels of 10 th Graders on MC Test..	109
Table 11: Proficiency Levels of 10 th graders on CR Test..	109
Table 12: Comparison of Proficiency Levels of 10 th graders on CAHSEE.....	110
Table 13: Pearson Correlation between MC and CR Scores..	111
Table 14: Pearson Correlations between MC and CR Questions by Gender.....	112
Table 15: Pearson Correlations between MC and CR questions by Strand.....	113
Table 16: T-test Results for Relationships between MC and CR scores by Gender.....	114
Table 17: T-test Results for Relationships between MC and CR Scores by Language.....	115
Table 18: Correlations between CST and CAHSEE.....	116
Table 19: Correlations between CST and CAHSEE for Gender and Language.....	117
Table 20: Correlations between MC and CR for Number Sense.....	133
Table 21: Correlations between MC and CR for Statistics and Probability.....	133
Table 22: Correlations between MC and CR for Algebra 1.....	134
Table 23: Correlations between MC and CR for Measurement and geometry.....	134
Table 24: Correlations between MC and CR for Algebra and Functions.....	135

CHAPTER I
INTRODUCTION

Overview

There are multiple ways to assess student learning in the field of mathematics. Methods range from standardized testing, using multiple choice and open-ended questions, to oral questioning and teacher-made examinations. This study focuses on the two formats used in state standardized tests: multiple choice (MC) and constructed response (CR).

Many questions can be raised about the potential differences between multiple-choice and free-response item formats. Multiple-choice (MC) tests are depicted as assessing simple factual recognition, and free-response or constructed-response (CR) tests are depicted as evaluating higher order thinking. A great deal of research has been devoted to comparing scores from multiple choice and constructed response tests (Bridgeman, 1992; Frederiksen, 1984; Ackerman & Smith, 1988). Many studies have also been dedicated to finding correlations between item format characteristics and race and gender. Some showed that there was a small advantage for men on multiple-choice items, and a small mean advantage for women on constructed response

items (Burton, 1996; Mazzeo & Schmitt, & Bleistein, 1991). Garner and Engelhardt (1999) investigated the gender differences in mathematics and found that women showed a statistically and consistent advantage over men on multiple-choice items in algebra. However, few studies have shed light on the performance of English Learners on free response compared to multiple-choice tests. There is a possibility that language ability might have a confounding effect on the scores for open-ended mathematics items and the fact that open-ended items are more likely to be omitted by examinees than multiple-choice items (Martinez, 1991).

The study aimed at finding relationships between mathematics scores in two formats - multiple-choice (MC) and constructed response (CR) items of the mock CAHSEE, differences in performance by gender and by language proficiency, as well as correlations between mock CAHSEE and CST scores. Statistical analyses were performed using correlations, descriptive statistics, and t-tests. Empirical data were also disaggregated and analyzed by gender, and language proficiency.

Theoretical Framework

The theoretical framework of the study is based on work by W. James Popham in educational measurement. In Popham's opinion, today's educators are increasingly caught up in a measurement-induced maelstrom focused on raising student scores on high-stakes tests. Standards-based standardized tests are in multiple-choice formats, with which teachers are more and more familiar. Due to intense pressure to raise students' scores, some teachers "design their instruction around actual items taken from a high-stakes test to teach toward *clone items* - items only slightly different from the test's actual items" (p.23). Because students are familiar with test content and format, they are trained to respond to questions by "recognizing" information, and may show mastery because they were strictly and specifically taught the content on the test.

The rationale of the study is to investigate the relationships between MC tests and their stem-equivalent constructed responses, allowing us to determine the degree to which student proficiency in one format relates to proficiency on the other.

Background

In the field of educational psychology, much of the literature suggests that item formats should be selected to reflect instructional intent, especially when trying to assess higher-level thinking. For instance, Haladyna (1997) writes that open-ended and performance items are more appropriate than selection items for measuring high-inference mental skills or abilities where we want the student to construct an answer. Rodriguez (2003) suggests that although multiple-choice tests provide greater sampling of the domain in a short time with a high level of reliability, the use of constructed response items allows greater depth of processes. One study found that teachers chose test formats according to the diverse achievement levels of their students (Fleming, Ross, Tollefson & Green, 1998). Those teachers assigned multiple-choice tests to low ability students and constructed response tests to students with higher cognitive abilities.

It is most generally assumed that multiple-choice tests do not adequately measure skills and cognitive abilities, and although they may measure some constructs, they may neglect others (Stenmark, 1989). Each person has

an individual profile of characteristics, abilities and challenges that result from learning and development. These are manifested as individual differences in intelligence, creativity, cognitive style, motivation, natures and the capacity to process information, communicate, and relate to others.

Advantages and Disadvantages of MC and CR tests

Both multiple-choice and constructed response items have advantages and disadvantages. Some of the advantages of MC items are that they are machine gradable, therefore increasing scoring accuracy (Holder & Mills, 2001); they are particularly useful in large-scale evaluation projects. They facilitate timely feedback for test takers in classes (Delgado & Prieto, 2003); and they enable instructors to ask a large number of questions on a wider range of subject materials (Becker & Johnston, 1999), therefore a wider variety of abilities can be measured. Other advantages are:

- Student difficulties can be diagnosed by analyzing incorrect responses.
- It is possible to vary the questions' level of difficulty.
- They are economical.

Some of their disadvantages are:

- They may not accurately measure student ability, since it may be assumed that they are guessing (Stenmark, 1989).
- Students are not able to synthesize content of any sort (Popham, 2010).
- They have an inability to tap higher order thinking skills.
- It takes a lot of time to construct a good MC test.
- The test is not useful in measuring the ability to organize and present ideas (Popham, 2010).

Some of the advantages of constructed response items are that results are reported in words, diagrams or graphs (Stenmark, 1989); and they give students an opportunity to show their prowess at carrying out a carefully reasoned analysis of the problem (Popham, 2010). One major advantage is that responses are less affected by guessing, and clues about students' thought processes can be provided. A few of the disadvantages of CR questions are that they contain relatively few questions, which in some cases prevents adequate sampling of the subject matter (Powell & Gillespie, 1990). They are costly, and there are potential inaccuracies associated with their scoring.

Standardized Tests and Assessment

Standardized tests are designed to assess student understanding of the content. They are formative and

summative criterion-referenced tests that measure how well a person has learned a specific body of knowledge and skills.

A variation of criterion-referenced testing is "standards-based assessment". All states and districts have adopted content standards (or curriculum frameworks), which describe what students should know to reach the basic, proficient, or advanced levels in the subject area.

Testwiseness and guessing

Testwiseness is any skill, which allows a student to choose the correct answer on an item without knowing the correct answer. Students who are testwise look for mistakes in test construction, make guesses based on teacher tendencies, and search for any unintentional clues that can be found in a test. This is an issue of validity because the score on a test should be a reflection of the level of the trait that the test is designed to measure (knowledge, skill), not a reflection of a general ability to do well on poorly made tests.

It is important to distinguish between random guessing and an educated guess. Good tests are designed to protect against random guessing. An educated guess is not as harmful to the validity of a test because it indicates that the student has some knowledge of the content and has

narrowed down the possibilities to the most reasonable alternative (Cronbach, 1998).

Reliability, Validity and Bias

Test reliability refers to the degree to which a test is consistent and stable in measuring what is intended to measure. It must be consistent within itself and across time.

Test validity refers to the degree to which the test actually measures what it claims to measure. It is the extent to which inferences, conclusions, and decisions made on the basis of test scores are appropriate and meaningful.

The presence of bias invalidates score inferences about target constructs that affect student performance differently across groups; constructs related to gender, race, ethnicity, linguistic background, and low socio-economic status (Lam, 1995). For example, the ability to read and understand written problems is a biasing factor in measuring mathematics skills because it is irrelevant to mathematics skills and it can affect Limited English Proficient students' performance differently on a math test (Stenmark, 1989).

A good assessment has both validity and reliability. In practice, however, an assessment is rarely valid or

reliable. In the field of educational testing, there will often be trade-offs between validity and reliability.

Significance Of The Topic

A review of the California State Department of Education's report on open-ended questions, *A Question of Thinking*, shows that most students lack opportunities to express mathematical ideas in writing, with fewer than 25% able to write completely about any of the problems given (Stenmark, 1989). Part of effective instruction is giving students opportunities to explain their thinking in writing, using proofs, multiple steps, organizers and written sentences.

Historically, there wasn't an emphasis on communication in the math classroom, but we now know that in order to learn mathematics, students must learn to communicate mathematically (NCTM 2000). This means listening, speaking, reading, and interpreting. It means explaining how a problem is solved, and explaining the problem and its solution using a variety of representations: words, symbols, graphs, charts, visuals, models, and manipulatives (Leiva, 1995).

The Principles and Standards of the National Council of Teachers of Mathematics (2000) include a communication

standard for school mathematics. Specifically, the standard states that instructional programs from kindergarten through grade 12 should enable students to:

- ❖ Organize and consolidate their mathematical thinking through communication.
- ❖ Communicate their mathematical thinking coherently and clearly to peers, teachers and others.
- ❖ Analyze and evaluate the mathematical thinking and strategies of others.
- ❖ Use the language of mathematics to express mathematical ideas precisely (p.60).

The more lessons focused on teaching conceptual understanding and problem solving, reading comprehension, and writing composition, the more likely the students were to demonstrate proficiency in all these areas (Knapp, Adelma, Marder, McCollum, Needles & Padilla, 1995).

The district where the research is conducted is plagued by dismal math scores on the California Standards Test. In four of the five comprehensive high schools, eighty percent of the students are scoring below and far below basic in mathematics, with under ten percent of students scoring in the advanced categories (California Department of Education, 2009).

Research Questions

This study attempts to find out if the students, as a group and by subgroups such as gender and English Language Learners, perform similarly on MC math tests and their stem-equivalent constructed response items.

Specifically, in this research, the following questions are being asked:

- 1) What is the relationship between the percents of students' correct answers on the multiple-choice format and correct answers on the stem-equivalent constructed responses? What are the differences by gender and language?
- 2) What is the relationship between students' math scores on the multiple-choice standardized mock CAHSEE test and their scores on stem-equivalent constructed responses?
- 3) Are there gender differences between the students' scores on the mock CAHSEE multiple-choice questions? Are there gender differences between students' scores on the stem-equivalent constructed responses?
- 4) Are there differences for English Learners (EL) between their scores on the multiple-choice questions and their stem-equivalent constructed responses? Are there differences for English Only (EO) students between their scores on multiple-choice questions and their stem-equivalent constructed responses?

5) What is the relationship between the students' mathematics California Standards Test and their scores on the multiple-choice?

6) What is the relationship between the students' CST scores and their scores on the constructed response tests on the mock CAHSEE?

Definition Of Terms

Multiple choice or selected response items (MC):

Multiple-choice items consist of a stem and a set of options. The *stem* is the beginning part of the item that presents the item as a problem to be solved, a question asked of the respondent, or an incomplete statement to be completed, as well as any other relevant information. The options are the possible answers that the examinee can choose from, with the correct answer called the *key* and the incorrect answers called *distractors*. Only one answer can be keyed as correct.

Constructed response, or open-ended response or free response (CR): A constructed response is a student response to a specific prompt or question given in the context of a test. It requires students to use creativity, organization skills, and logic to develop an answer. Most commonly, a constructed response takes the form of an essay response or a short-answer response.

Stem-equivalent: Multiple-choice and constructed response questions will have the same *stem*, which is basically a math question or a problem to be solved. For example, if a student is asked a question about finding the perimeter of a figure, the MC test will provide the optional answers, and the CR test will ask the same question and the student will have to show the solving process.

Standardized testing: Tests are called *standardized* when all students answer the same questions under similar conditions and their responses are scored in the same way. They include norm-referenced tests as well as criterion-referenced or standards-based exams.

The CAHSEE: The California High School Exit Exam (CAHSEE) is a requirement for high school graduation in the state of California, created by the California Department of Education to improve the academic performance of California high school students, and especially of high school graduates, in the areas of reading, writing, and mathematics; public school students must pass the exam before they can receive a high school diploma, regardless of any other graduation requirements.

Methodology

Research Design

A number of statistical analyses were used. Correlations were run to determine relationships between scores on both testing formats (MC and CR), as well as between these scores and those on the California Standards Test in Mathematics. Frequency tables were run to investigate percentages of students scoring at various levels of proficiency on both formats. T-tests were also performed using gender, and language (English Learners versus English Only).

Sampling

The sample consisted of 737 students enrolled as freshmen (n= 394) and sophomores (n= 343) in algebra 1, algebra 2 and geometry at a comprehensive high school in the Pomona Unified School District. The majority of the students were Latinos, but there were also Asian students of different ethnic backgrounds, African American students, and some white students. The ethnicity variable was initially considered but the comparably insignificant percentage of non-Latinos (9%) caused it to be discarded.

Instrumentation

The instrument is the Mock CAHSEE in mathematics. It is a test designed by the district to help the students

familiarize themselves with the content before taking the actual CAHSEE, and it is aimed at assessing student knowledge in order to plan for intervention and remediation by the time they take the CAHSEE. All of the 35 questions on the tests cover the mathematics standards required to pass the CAHSEE. Eleven questions are related to Number Sense, four are related to Statistics and Probability, four are related to Algebra and Functions, six to Algebra 1, and ten to Measurement and Geometry.

Procedures

Thirty-five questions were selected from the Mock CAHSEE math booklet (2008 edition) in such a manner that they reflected different standards from the strands of Number Sense, Statistics and Probability, Algebra and Functions, Algebra One, and Measurement and Geometry. It is customary at this particular school to administer the Mock CAHSEE to ninth graders on the day that the tenth graders are taking the actual CAHSEE. The school is on a special schedule because the test is administered all day, from 8 a.m. to 1:30 p.m. Twelve teachers administered the test to 394 Freshmen, who were given the test in constructed response format first, then in multiple-choice format later in the day after a thirty-minute lunch break from 10:30 to 11:00 a.m.

The tenth graders were given the test before the 9th graders, in their math classes two weeks before they were to take the CAHSEE. All math teachers agreed to give the multiple-choice format test first on the same day, and waited to give the constructed response test the following week over a period of two days.

Protection of human subjects

All scantrons and constructed response tests had student ID numbers written on them to protect the identity of the students. The students were previously handed a consent form to be signed by their parents, and an assent form to be signed by them agreeing to take the test willingly. They were all aware that it was not just per school policy that the test was given, but that their scores would be evaluated for the purpose of the study. The results of the study will only be released to their teachers or administrator of the school as was previously agreed upon and approved before the launch of the experiment.

Scoring rubric

The California Mathematics Council rubric is called a general, or holistic, rubric and is used on national or state assessments that must take into account a broad range of mathematical tasks and students. It is aimed at

assigning an overall score rather than a score for particular processes. This type of rubric is appropriate for assessments that are more summative, such as major tests or examinations (Kulm, 1994). "The descriptions of each score are precise enough so that in a short time, teachers can be trained to use the scoring scale with high levels of agreement and reliability" (p.88).

Summary

An extensive review of the literature describing the various findings on the different testing formats is discussed in Chapter Two. Issues such as the advantages and disadvantages of MC and CR tests, as well as reliability and validity issues in writing those tests are also included. Chapter Three explains the methodology used in the study, the data set, the procedures and the instrumentation.

Descriptive statistics, correlations and t-tests are presented in Chapter Four. Results from this analysis provide insight into the results of various formats with different groups of students. The implications of the study findings are discussed in detail in Chapter Five.

CHAPTER II

LITERATURE REVIEW

Overview

Testing formats have their advantages and disadvantages. Previous studies have lauded the effectiveness of some formats in assessing student learning, while denigrating other formats for their poor assessment quality. In mathematics, notably, it is most important to discern and evaluate the effectiveness, or lack thereof, of the testing formats in an effort to select the best method of assessing student content knowledge.

Constructed Response Tests

Advantages. The California Mathematics Council (CMC) has been a leader in stressing the use of open-ended questions as a technique of alternative assessment. Open-ended questions provide insights into the misconceptions of students and allow the teacher to evaluate the various techniques they use. They also determine if students can “clarify their own thinking, make generalizations, recognize key points in the problem, and organize and interpret information” (Kulm, p.42).

Constructed response tests reduce measurement error by eliminating random guessing. Second, they eliminate unintended corrective feedback that is inherent with MC items (Bridgeman, 1992). Bridgeman (1992) found that 81% of the students reported working backwards to solve problems. For example, an algebra problem such as $2(x+4)=38-x$ becomes a much simpler arithmetic problem if the examinee can just substitute the possible values of x given in the answer choices until the correct value is found.

A constructed-response test allows us to watch a student marshal evidence, arrange arguments, and take purposeful action to address the problem (Wiggins, 1989). Rather than rely on right or wrong answers and unfair "distractors", authentic tests identify strengths, which may even be hidden (Wiggins, 1989). They assess dynamic cognitive processes (Bennett, Ward, Rock, & Lahart, 1990), identifying students' misconceptions in diagnostic testing (Birenbaum & Tatsuoka, 1987), and communicating to teachers and students the importance of practicing these real-world tasks (Sebrechts, Bennett, & Rock, 1991).

Haladyna (1997) writes that open-ended and performance items are more appropriate than selection items for measuring high-inference mental skills or abilities and some physical skills and abilities where you want the

student to construct an answer. In order to assess higher order thinking, they argue that performance assessments are a more appropriate item type than selection items because they require students to construct new knowledge, which is essential to effective learning (Marzano, Pickering, & McTighe, 1993).

The shift from an emphasis of producing correct answers to the expectation that students think and communicate is a major one for many students and teachers (Kulm, 1994). Even though the answer may not be correct, the reasoning and mathematical processes can earn high marks.

Open-ended problems must be provided to all students, even the most able ones, if we want them to develop solving strategies. The process and strategies themselves must be the objects of assessment and evaluation (Kulm, p.26).

Some of the advantages of constructed response items are that results are reported in words, diagrams or graphs (Stenmark, 1989); and they give students an opportunity to show their prowess at carrying out a carefully reasoned analysis of the problem (Popham, 2010). One major advantage is that responses are less affected by guessing, and clues about students' thought processes can be provided.

Open-ended questions send out a message to students about the nature of math (Brahier, 2001). Students "learn" that mathematics transcends "right" and "wrong" answers (p.22). Marzano et al. (2001) stress that explaining their thinking helps students to enhance their understanding of the experimental inquiry process and their use of the steps involved. Also, the *range* of cognitions - such as knowledge, procedures, images and skills - that can be elicited by CR items is greater than the range of MC items (Martinez, 1999).

Disadvantages. There are many things to consider when choosing between constructed-response and selected-response tests. Constructed-response tests are much more difficult to grade, even though they are relatively easy to prepare. A considerable amount of time must be spent in creating clear criteria, such as scoring rubrics, for assessing the answers. One of the most evident disadvantages is the time-consuming nature of scoring those tests. The scoring of constructed-response test items involves at least some subjectivity, even when criteria have been carefully established (Powell & Gillespie, 1990; Brahier, 2001). Another disadvantage is that these tests contain relatively few questions, which in some cases prevents adequate sampling on the subject matter.

Test anxiety may have a debilitating effect on scores. Research by Crocker and Schmitt (1987) found that the negative effects of test anxiety on scores were moderate on MC questions but severe on the constructed response items. The prospect of having to provide an explanation can induce anxiety to the point that it interferes with cognition, therefore reducing the ability of the test taker to express proficiency (Powers, 1988). Popham (2008) suggests that if there were too few items, odds were greater that the teacher would "draw an invalid inference from the performance data, concluding erroneously that students have or have not mastered the building block to an acceptable degree" (p.58).

Open-ended questions may not align with instructional techniques (Brahier, 2001). If students are not often asked these types of questions in the classroom, it may be unrealistic to expect them to answer open-ended questions on a more formal assessment (p.22). As Kulm (1994) points out, most students have not been required or requested to write or give verbal explanations of problem-solving processes. "The idea of an assessment or grade based on anything except the correct answer is quite foreign" (p.39).

Multiple-Choice Tests

Advantages. Some of the advantages of MC items are that they are machine gradable, therefore they increase scoring accuracy (Holder & Mills, 2001), and they are particularly useful in large-scale evaluation projects (Dufresne, Leonard & Gerace, 2002). They facilitate timely feedback for test takers in class (Delgado & Prieto, 2003); and they enable teachers to ask a large number of questions on a wide range of subjects (Becker & Johnston, 1999), therefore a wider range of abilities can be measured. Student difficulties can be diagnosed by analyzing incorrect responses, and it is possible to vary the questions' difficulty level (Simkin & Kuechler, 2005). Roediger and Marsh (2005) postulate that in addition to being easy to score, multiple-choice tests generally improve student performance on later tests, referring to that as the *testing effect*. There is a perceived objectivity in the grading process (Wainer & Thissen, 1993); they help students avoid losing points for poor spelling or poor writing ability (Zeidner, 1987); students find it easier to prepare for those tests (Scouller, 1998); they reduce student anxiety (Snow, 1993); teachers may choose to write multiple versions of the same MC test to thwart cheating

(Kreig & Uyar, 2001); students can eliminate unlikely choices and ultimately increase their probability of picking the right answer (Bridgeman, 1992).

Multiple-choice items are amenable to item analysis, which enables the teacher to improve the item by replacing distractors that are not functioning properly. In addition, the distractors chosen by the student may be used to diagnose misconceptions of the student or weaknesses in the teacher's instruction (Burton et al., 1991).

Disadvantages. Some of the disadvantages are that they may not accurately measure student ability, since it may be assumed that they are guessing (Stenmark, 1989); students are not able to synthesize content of any sort (Popham, 2010); and they have an inability to tap higher order thinking skills. It takes a lot of time to construct a good MC test; the test is not useful in measuring the ability to organize and present ideas (Popham, 2010). The format makes it easy for students to guess rather than to think through the problem.

MC items have an inability to tap higher order thinking and allows for a higher probability of guessing correctly which causes lower reliabilities in the test for lower ability students (Cronbach, 1988). By design, MC items severely constrain the behavior of examinees.

Consequently, some aspects of proficiency that require complex performance are beyond the reach of the MC format (Messick, 1993). If a test consists entirely and exclusively of MC items, it raises the possibility of construct under-representation and the validity of the assessment will suffer because the test will fail to assess the cognitive processes that help identify the main construct (Messick, 1995).

Webb (1997) argues that multiple-choice tests inherently favor some students over others, so alternative forms of assessment are required to achieve fair measures of student performance. Hambleton & Murphy (1992) concluded that multiple-choice tests foster a one-right-answer mentality, they narrow the curriculum, they focus on discrete skills, and they under-represent the performance of lower SES examinees. Martinez (1991) argues that language ability might have a confounding effect on the scores for open-ended mathematics items and that open-ended items are more likely to be omitted by the examinee than multiple-choice items.

Test takers are exposed to numerous incorrect answers, many of which are constructed so as to appear to be correct. Roediger (2005) found that students tended to remember these incorrect lures as to be correct when asked

about them later, suggesting that students actually learn the wrong things as part of the testing process. A related disadvantage is that students receive corrective feedback whenever their own answer does not appear as one of the available alternatives, a prompt to *reconsider* the question and correct their mistake that would not be present in an open-ended assessment (Bridgeman, 1992). Some students react to the availability of the possible answer by working backwards to answer the question, particularly on quantitative problems. Students expecting a multiple-choice test, relative to an essay test, spend less time studying for the test (Kulhavey, Dyer, & Silver, 1975) and they take notes on different materials than do students expecting an essay exam (Rickards & Friedman, 1978).

According to the NCTM (1991), although the commonly used MC format may yield important data, it can have a negative impact on how students are taught and evaluated at the school level because: a) Student scores are generated solely on the basis of right and wrong answers with no consideration or credit given to students' strategies, b) Routine timing measures how quickly students can respond but not necessarily how well they think - some students may be excellent mathematicians but may not be fast (p.22), and

c) Mathematics tools such as calculators and measurement devices are not permitted (p.8).

MC Items versus CR Items

How they differ. Martinez (1999) hypothesized that MC and CR item formats differ not only in their cognitive demand but also in the range of cognitions they can elicit. And even though the distinction between them is useful, it could be misleading. In his meta-analysis of research on test item formats, Martinez (1999) discusses research pertaining to the complexity of both MC and CR formats. Haladyna (1994) proposed that there was considerable variety within the MC format, partly in how items are structured and in the cognition they evoke. He further asserts that MC items *can* be written to elicit complex cognitions, such as understanding, prediction, evaluation, and problem solving. In other words, it is possible for the MC items to tap complex performances and for CR items to tap basic processes such as recall. And even when MC items evoke recall, the retrieval of information from long-term memory may require complex search strategies to access memories from various learning episodes (Nuthall & Alton-Lee, 1995). Messick (1995), however, warns that even though MC questions can be designed to elicit complex thought

processes, it does not mean, however, that the full range of complex thought represented in constructed responses can be captured by MC items.

Many studies have found that student scores on open-ended questions were so closely related to their scores on multiple-choice tests as to suggest that both types of questions were measuring the same things (Bridgeman, 1992; Lukhele, Thissen, & Wainer, 1994; Walstad & Becker, 1994), suggesting that the difficult to administer open-ended questions might not be worth the extra effort because multiple-choice alone could be used to assess the learning. Popham (1978, pp. 44-45) states that for measuring knowledge of factual information, the selected-response test is more efficient. This type of test is also useful when a high degree of specificity is needed, such as tests designed to see if re-teaching of facts is necessary. However, for measuring originality, the ability to synthesize ideas, write effectively, or solve problems, constructed-response tests are obviously better.

In an experiment led by Fleming (1998), it was found that teachers assigned tests of different formats based on students' cognitive abilities. Low ability students were given MC tests and high ability students were given essay type or constructed-response test items. They concluded

that teachers judged essay questions to be more difficult than multiple-choice items, and they evaluated items that measured higher order thinking skills to be more difficult than items assessing application or memory skills.

Format preference. In a study by Hamilton (1994) high school students enrolled in geometry, algebra 2 and algebra 1 were given a math test with multiple-choice and constructed-response formats in counterbalanced order. After taking the tests, students were interviewed to determine which format was preferred and why. Eighty percent of students found MC to be easier. Several students also recognized that the probability of answering an item correctly when they did not know the answer was much greater for MC than CR. Over fifty percent of the students who preferred the CR test reported that they liked the challenge it presented. Although the majority of students preferred the MC test, a very small percentage said that it was a better indicator of what they knew.

Parmenter (2009) reflects that the literature tends to favor multiple-choice over constructed-response as far as validity and reliability were concerned. For example, Bridgeman (1992) suggested that although multiple-choice is less reliable on a question by question basis due to guessing, the fact that multiple-choice questions take less

time to answer and grade would allow an exam made up entirely of multiple-choice to contain more questions and therefore be more reliable than an exam containing fewer open-ended questions. It is generally assumed that correct answers to MC items can be guessed at more readily than CR items, thus MC tests are less difficult, less discriminating and less reliable than CR tests of the same content. In addition, having multiple answers - one of which is the correct one - may alert the examinee who makes a mistake in the computation and ends up with an answer that is not on the list of choices, to check and /or redo the computation. However, these expectations are not supported by findings of empirical research (Traub and McRury, 1990).

Traub and McRury (1990) report that students have more positive attitudes towards multiple choice tests in comparison to free response tests because they think that these tests are easier to prepare for, easier to take, and thus will bring in relatively higher scores. In the study by Ben-Chaim and Zoller (1997), the examination format preferences of secondary school students were assessed by a questionnaire and structured interviews. Their findings suggested that students preferred written, unlimited time examinations and those in which the use of supporting

material was permitted. Assessment formats, which reduce stress will, according to these authors, increase the chance of success and students vastly prefer examinations which emphasize understanding rather than rote learning.

Martinez (1999), however, describes the students' preferences of CR formats as just a "perception". Their opinions did not constitute reliable evidence that MC items tapped lower-level cognitive processes. Birenbaum (1997) found that differences in assessment preferences correlated with differences in learning strategies. Moreover, Birenbaum and Feldman (1998) discovered that students with a deep study approach tended to prefer essay type questions, while students with a surface study approach tended to prefer multiple-choice formats. Students with high test anxiety had more favorable attitudes toward multiple-choice questions while those with low test anxiety tended to prefer open ended formats (Birenbaum, 1997).

Scouller (1998) investigated the relationships between students' learning approaches, preferences, perceptions, and performance outcomes in two assessment contexts: a multiple-choice question examination requiring knowledge across the whole course, and assignment essays requiring in-depth study of a limited area of knowledge. The results indicated that if students preferred essays, then they

would do better on the essay items than if they preferred multiple-choice questions.

Study skills and performance. A review of the California State Department of Education's report on open-ended questions, *A Question of Thinking*, showed that most students lacked opportunities to express mathematical ideas in writing, with fewer than 25% able to write completely about any of the problems given (Stenmark, 1989). According to NCTM (1991), it is the task that requires students to construct their own responses that more closely models real work and prepares students for life outside school. Tests that emphasize narrow recall will not effectively prepare students for a world that demanded thinking and communication. There is evidence that students study differently depending on the type of test they anticipate and this alters the nature and quality of student learning. Studies are mixed in their detection of anticipation effects; however a majority of studies have found that response formats make a difference in anticipatory learning and that the expectation of CR tests favors concept learning while the anticipation of MC tests favors detail memorization (Martinez, 1999; Traub & McRury, 1990). Douglas Reeves, chairman and founder of the Center for Performance Assessment and the International Center for

Educational Accountability, has said that "even if the state test is dominated by lower-level thinking skills and questions are posed in a multiple-choice format, the best preparation for such tests is not mindless testing drills, but extensive student writing, accompanied by thinking, analysis, and reasoning" (2004, p. 92).

Testwiseness. Testwiseness is any skill, which allows a student to choose the correct answer on an item without knowing the correct answer. Students who are testwise look for mistakes in test construction, make guesses based on teacher tendencies, and search for any unintentional clues that can be found in a test. Millman, Bishop and Ebel (1965, in McPhail, 1981) known for their theoretical work on testwiseness proclaim that "testwiseness is defined as a subject's capacity to utilize the characteristics and format of the test and/or the test taking situation to receive a high score. Testwiseness is logically independent of the examinee's knowledge of the subject matter for which items are supposedly measured". (McPhail, 1981, p.707).

A number of researchers have investigated the belief that the results of MC tests can be influenced by "testwiseness" (Simkin & Kuechler, 2005). The most common technique is to eliminate one or more MC answers based on

only a partial understanding of the knowledge being tested and thus generate misleadingly high test scores. Studies by Rogers and Hartley (1999) and Zimmerman and Williams (2003) both corroborate the influence of testwiseness on MC examinations. Researchers have found that testwiseness skills introduced additional variance into examination scores (Fagley, 1987), and that there was a positive association between testwiseness skills and classroom examination performance (Fagley, Miller, and Downing, 1990). Teaching testwiseness would improve the validity of test results, were likely to strengthen critical thinking, and provided equal education, employment and opportunity for minorities (McPhail, 1981). There are two ways of learning testwiseness: associative learning and problem solving. Associative learning means learning from being told and from practice and drill. In problem solving, students search for a pattern; they are presented with evidence and are asked to investigate the data and draw conclusions (McPhail, 1981).

It is also beneficial to raise English Language Learners' awareness of the typical discourse and formats of standardized tests. ELLs may not be familiar with the kind of language that is used in tests, including many

predictable patterns and phrases. It may also be beneficial to teach test-taking skills (e.g., how to approach a multiple-choice question, how to locate the main idea in a reading passage) to help prepare ELLs for specific types of test items they may encounter. Armed with a variety of test-taking skills and strategies, ELLs may be empowered to demonstrate their knowledge on a test, rather than being intimidated by unfamiliar terms and formats (McPhail, 1981).

Guessing. Differences among students on variables that affect the amount of guessing have been identified as a source of error on multiple-choice tests (Cronbach, 1980). Guessing on a multiple-choice item may be categorized as random (among all choices), or informed (where some wrong choices are eliminated (Frary, Cross & Lowry, 1977)). Most researchers agree that the influence of blind guessing on the scores of a test diminishes as the length of a test and the number of options per item increases (Ebel & Frisbie, 1991). The guessing factor reduces the reliability of multiple-choice item scores somewhat, but increasing the number of items on the test offsets this reduction in reliability. For example, if the test includes a section with only two multiple-choice items of 4 alternatives each (a b c d), one can expect 1 out of 16 of your students to

correctly answer both items by guessing blindly. On the other hand if a section has 15 multiple-choice items of 4 alternatives each, you can expect only 1 out of 8,670 of your students to score 70% or more on that section by guessing blindly (Burton et al, 1991).

Gender

Research studies have shown that male/female differences on constructed-response questions often do not parallel the male/female differences on the multiple-choice questions in the same subject (Mazzeo, Schmitt, & Bleistein, 1992). Typically, when women and men perform equally well on the multiple-choice questions, the women outperform the men on the constructed-response questions. When women and men perform equally well on the constructed-response questions, the men outperform women on the multiple-choice questions. The differences occur even though the multiple-choice scores and the constructed-response scores tend to agree strongly within each group. In academic subjects, there is usually a strong tendency for the students who are stronger in the skills measured by the multiple-choice questions to be stronger in the skills measured by the constructed-response questions. But if all students improve in the skills tested by the CR questions,

their performance on the MC questions may not reflect that improvement (Livingston, 2009).

Learning Strategies. Kimball (1989) hypothesized that gender-related differences in performance are the result of different approaches to learning mathematics. Gallagher (1992) found that most of the items favoring men required insightful strategies, whereas all the items favoring women required standard algorithmic strategies.

Format preferences. In a study done by DeMars (1997), scores from mathematics and science sections of pilot forms of a high school proficiency test were examined for evidence of an interaction between gender and response format (MC or CR). When students of all ability levels were considered, the interaction was small in science and non-existent in math. When only the highest ability students were considered, male students scored higher on the multiple-choice section, whereas female students either scored higher on the constructed-response section or the degree to which the male students scored higher was less on the constructed-response section. Correlations between the formats were high and did not vary by gender.

Beller and Gafni (2000) gave an overview of several studies, which analyzed the students' preferences for assessment formats, their scores on the different formats,

and the influence of gender differences. In a range of studies, they found some consistent conclusions suggesting that, if gender differences are found (which was not always the case), female students preferred essay formats, and male students showed a slight preference for multiple-choice formats. Furthermore, male students scored better on multiple-choice questions than female students and female students scored better than male students on open-ended questions than on multiple choice questions (Ben-Shakhar and Sinai, 1991; DeMars, 1997).

MC and CR formats require different sets of skills, and these skills may differ for genders. An example is the influence of verbal fluency for writing tasks. Some studies have found that females have higher verbal fluency than males (Halpern, 1992). If this is true, these higher fluency skills may give females an advantage over males in CR tasks. Willingham and Cole (1997) reviewed national and state assessment results and concluded that writing often appeared to play a role in gender format score differences. The research they reviewed suggested writing skills and fluency differences as possible factors in the female advantage on CR tasks. They also reported that requested discussion and explanation of responses consistently showed female advantages. Clements and Ballista (1992) suggested

that males and females differ on preferred solution strategies with more females choosing verbal strategies and more males choosing non-verbal strategies.

The age factor. In a meta-analysis performed by Hyde, Fennema, and Lamon (1990) on gender differences in mathematics performance, they found that overall differences in mathematics performance were not apparent in early childhood, but that they appeared in adolescence and usually favored boys in tasks involving high cognitive complexity, such as problem-solving, and favored girls in tasks of less complexity, such as computation. In addition, there was a slight female superiority in performance in the elementary and middle school years. A moderate male superiority emerged in the high school years. Females were superior in computation in elementary and middle school, and the difference was essentially zero in the high school years. The gender difference was essentially zero for understanding of mathematical concepts at all ages for which data was available. It was in problem solving that dramatic age trends emerged. The gender difference in problem solving favored females slightly in the elementary and middle school years, but in the high school and college years, there was a moderate effect size, favoring males. It was assumed that this occurred because in high school and

college, students were permitted to select their own courses, and females chose fewer mathematics courses than did males (Meece, 1992). Differences in course selection appeared to account for some but not all of the gender difference in performance on standardized tests in the high school and college years (Kimball, 1989).

Ethnicity and Language

According to the recently published Guidelines for the Assessment of English Language Learners, by the Educational Testing Service (2009), English Language Learners (ELLs) represent one in nine students in U.S classrooms from pre-Kindergarten through 12th grade, but most are concentrated in the lower grades. Eighty percent are native speakers of Spanish, and about five percent are of Asian descent. English Language Learners are concentrated in six states- California, Arizona, Texas, New York, Florida and Illinois. In California, more than 25% of the students in grades pre-K-12 are ELLS.

ELLs vary greatly as individuals. Therefore, there is no particular response format that is most advantageous for all. If the multiple-choice format is decided upon, large amounts of texts make it less likely that they will understand what is being asked of them (Martiniello, 2008).

If the constructed-response format is selected to assess their knowledge, the examiner might consider including tasks that allow examinees to respond, not in long, wordy sentences, but in diagrams or other visual representations (Snow, 2000). It may be challenging for students learning English to show what they know and can do in mathematics if the test items that assess this knowledge also test their English language skills. The complexity of the language in a math test item may interfere with the ability of ELLs to demonstrate their understanding of math concepts on achievement tests (Abedi, 2002). Mathematics test items can be reworded to minimize their language load without altering the content assessment (Abedi, 2002).

Low scores on a standardized test may mean nothing more than that a learner has not yet mastered enough English to demonstrate his or her content knowledge and skills on a test. Multiple assessments, including some performance-based or alternative assessments that mirror what students are learning in class, will paint a much more accurate picture of students' knowledge, skills, and progress than any single test score can indicate (Coltrane, 2002).

Accommodations. Using Mathematics test items from the National Assessment of Education Progress (NAEP), Abedi et al (2002) employed accommodation strategies (modified English, use of dictionary, extra time) and the results indicated that ELL students scored, on average, 5 points lower than non-ELL students on a 35-item math test. Also, students who were better readers achieved higher math scores. In an earlier study using the 1990 NAEP Mathematics Assessment, it was found that members of some ethnicities were less likely to respond to open-ended items than were students in other groups. This finding suggests that the experiences students bring to the testing situation may interact with test format to influence their performance, and that elimination of the multiple-choice format may increase, rather than reduce, achievement gaps (Myerberg, 1996).

Bronwyn Coltrane of the Center for Applied Linguistics advocates teaching ELLs the discourse of tests and test-taking skills: "It is. . . beneficial to raise ELLs' awareness of the typical discourse and formats of standardized tests. ELLs may not be familiar with the kind of language that is used in tests, including many predictable patterns and phrases. It may also be beneficial to teach test-taking skills (e.g., how to approach a

multiple-choice question, how to locate the main idea of a reading passage) to help prepare ELLs for specific types of test items they may encounter. Armed with a variety of test-taking skills and strategies, ELLs may be empowered to demonstrate their knowledge on a test, rather than being intimidated by unfamiliar terms and formats". This preparation in how to approach test questions and answer sheets is especially important for ELLs who are recent immigrants. Even those who have some proficiency in English may never have been exposed to the format of U.S. standardized testing.

Scoring Rubrics

Scoring constructed-response items written by ELLs may present additional challenges. Two ways in which ELLs' constructed responses differ are differences due to language background and in the style of the response (Abedi & Lord, 2001). For example, if they have to use sentences to write a proof, one must overlook errors in grammar and syntax, and focus on the *content* knowledge and the *range* of that knowledge. Also, arithmetic operations are learned differently in other countries. To name a few, the conventions for long division are different, and decimals are expressed as commas in Europe and Asia.

Formatting

Formatting is important for students whose processing strategies and decoding efforts result in literacy and language challenges (Abedi, 2002). Some critics suggest that, for ELLs, the most humane approach is to focus almost exclusively on the reduction of language in the text. In mathematics, for instance, asking to solve " $3x + 5x$ " would be more fitting and less confusing than asking to solve "the sum of three times a number and five times that same number". Although it may seem like English Language Learners may fare better on multiple-choice tests because they are not obligated to express their reasoning in writing - which may prove to be weak - testing them largely or exclusively on multiple choice tests may mask their real abilities.

Empirically, Kopriva and Lowrey (1994) found that a large percentage of ELLs in California said they would rather have an open-answer format as compared with multiple-choice format for providing their responses. They said that the CR format provided them with the chance to explain what they know. It is further recommended then that CR items be used to allow for different approaches to demonstrating mastery, such as charts, diagrams and pictures.

Edwards and his colleagues (2007) investigated subgroup differences on a multiple-choice and constructed-response test of scholastic achievement in a sample of African American and White students. Although both groups had lower mean scores on the constructed-response test, the results showed a 39% reduction in subgroup differences compared with the multiple-choice test. That proved that African Americans had more favorable perceptions on the constructed-response tests. The authors concluded that integrating constructed-response items would be a viable alternative for minimizing subgroup differences on high-stakes testing.

Validity

Many researchers and practitioners believe that standards-based reform and high-stakes testing will have the greatest impact on Blacks, Latinos, English-language learners, students with disabilities, and low-SES students (Heubert, 2009). As beneficial as it may be to include ELLs in high-stakes tests, some complications arise concerning the validity and reliability of such tests for this group of learners (Coltrane, 1992). Educators must consider what is actually being assessed by any given test: Is the test measuring ELLs' academic knowledge and skills, or is it primarily a test of their language skills? When ELLs take

standardized tests, the results tend to reflect their English language proficiency and may not accurately assess their content knowledge or skills (Menken, 2000), therefore weakening the test's validity for them. If ELLs are not able to demonstrate their knowledge due to the linguistic difficulty of a test, the test results will not be a valid reflection of what the students know and can do.

Popham (1999) hints that there are test questions that "may appear to be appropriate for assessing students' skills and knowledge, but in reality, there is a real presence of SES-linked content that gives an edge to children, whose parents are middle or upper class, are better off financially or have received a higher education" (p.59). Perhaps most importantly, educators must be cautious when interpreting the test results of ELLs. As with all learners, it is crucial to remember that one test cannot accurately reflect everything that a person has learned and is able to do. This point is particularly important if the validity and reliability of the test are questionable for ELLs, or if the students were not given appropriate testing accommodations. Similarly, high-stakes decisions should not be made regarding a program, school, or district with high numbers of ELLs based solely on test data. Such data may merely indicate that a school or

district has a high percentage of ELLs, and not be reflective of instructional quality or program effectiveness (Menken, 2000).

Guidelines for Writing Multiple-Choice Questions

From a teaching and learning point of view, question construction has to address specific criteria for good assessment (Earl, Land and Wise, 2000). The questions have to be a) reliable: they must produce consistent results, b) valid: the question must test what the student has been taught, c) useful: the assessment must help the student progress and reinforce the learning, d) fair: all students who take the assessment should have an equal chance of scoring full marks, and e) cost effective: questions must be efficient enough to produce the required results for the students and the institution in general.

Haladyna and Downing (1989) are recognized as major contributors to the research on multiple-choice testing. They devised guidelines for procedural and content item writing, as well as stem construction and option and distractor development. They advise the following:

1. Avoid the complex multiple-choice (Type K) format. (e.g., A and D, A and C, All the above, None of the Above, A, B and C, etc.).
2. Minimize examinee reading time in phrasing each item.

3. Avoid trick items, which mislead or deceive examinees into answering incorrectly.
4. Base each item on an educational or instructional objective.
5. Keep the vocabulary consistent with the examinees' level of understanding.
6. Use multiple-choice to measure higher-level thinking.
7. Test for important or significant material; avoid trivial material.
8. State the stem in question form or completion form (*note: recent research findings favor question form over completion*).
9. Ensure that the directions in the stem are clear, and that wording lets the examinee know exactly what is being asked.
10. Avoid window dressing (excessive verbiage) in the stem.
11. Word the stem positively; avoid negative phrasing.
12. Include the central idea and most of the phrasing in the stem.
13. Use as many options as are feasible; more options are desirable.
14. Place options in logical or numerical order.
15. Keep the length of the options fairly consistent.
16. Avoid, or use sparingly, the phrase "all of the above."

17. Avoid, or use sparingly, the phrase "none of the above."
18. Avoid the use of the phrase "I don't know."
19. Avoid distractors that can clue test-wise examinees; for example, avoid clang associations, absurd options, formal prompts, or semantic (overly specific or overly general) clues.
20. Avoid giving clues through the use of faulty grammatical construction.
21. Avoid specific determiners, such as "never" and "always."
22. Make sure there is one and only one correct option.
23. Use plausible distractors; avoid illogical distractors
24. Incorporate common errors of students in distractors.
25. Avoid technically phrased distractors.
26. Use familiar yet incorrect phrases as distractors.
27. Use true statements that do not correctly answer the item.

Guidelines for the Constructed-Response Items

There exist many references on how to construct valid constructed response items. General guidelines can be gleaned and summarized as follows:

1. Design CRs so that students are challenged to think and not just to provide memorized answers.

2. CRs should be very clear about what the students are to do. The stem should focus the students to the questions/tasks but not so narrowly that a students' response cannot be scored on all scoring levels.
3. Ask the student to "define, explain, or identify..."
4. Ask the student to "explain why...." Without the understanding component, a CR is only requiring a student to recall information.
5. Ask the student to "include details and specific examples to support your answer."
6. Do not use the verbs "discuss", "think about", "illustrate" or "consider". Use "explain," "justify," or "describe" instead.
7. Utilize Bloom's Taxonomy as you write your essay questions. Focus on the higher levels of the taxonomy, such as analysis, synthesis and evaluation. Bloom's Taxonomy provides sample ideas of what students should know and be able to do at each level of understanding; take these samples and turn them into essay questions.
8. Give your students clear guidelines for how to answer the essays. When you write your questions, think about how you want your students to answer them. Use this knowledge to develop a scoring rubric, and include it with the test. This way your students will have a guideline to use as they

write, and they will have a better chance of earning a good score on their essay tests.

Effective Mathematics Instruction

Multiple-choice tests were created initially for their practicality in saving time and money when given on a very large scale. In most recent years, testing deadlines, increasing data-driven accountability, and the very challenges of teaching English Learners, have gradually shifted many teachers' focus from teaching higher thinking skills to teaching to the test.

Popham (2000) introduced two perceptions worth considering when preparing students to take tests: (1) a test's items and (2) the knowledge and/or skills represented by those items. He claims that if a teacher directed instruction toward the body of knowledge and/or skills a test is supposed to represent, then we would applaud that teacher's efforts. This kind of instruction can be called teaching toward the knowledge and/or skills *represented* by a test. However, he adds, a teacher who either uses the test's *actual* items in classroom instructional activities or uses items so similar to the test's actual items as to be almost indistinguishable, is one who is remiss of his duties as an effective educator.

If teachers implemented efficient strategies and instilled motivational and cognitive activities in their instruction of mathematics, they would focus more on the students' level of mastery using various tools of assessment, and will feel more confident about the students being prepared to take any test regardless of the format.

Explicit Instruction

Given the current trend of teaching all students, including English Language Learners, it is important to find instructional approaches that adequately address the diverse needs of students. This is particularly challenging when it comes to mathematics instruction. Leading educators, researching instruction for students with diverse learning needs have continued to support an explicit teaching methodology for teaching mathematics (Carnine & Gersten, 1982).

Effective teaching is the orchestration of many skills into a coherent system that meets the need of a class. All the experts on effective teaching have discerned essential qualities in teachers that were instrumental in moving their students forward. As found in the literature, teachers who favored the direct instruction approach were most successful in inculcating meaning, comprehension, and skills in students who needed it the most.

Knowledge and mastery of the content

An effective teacher is an expert in the content he teaches (Haberman, 1995; Hirsch, 2007; Ravitch, 2000). He must be well educated, not just well trained in his field of discipline (Ravitch, 2000). Mastery of the content knowledge influences instructional practice: the more knowledgeable the teacher is in his mathematics field, the more confident he will feel in imparting information to his students, and finding answers to their questions. Excellent teachers know their subject area and possess a flexible repertoire of pedagogical strategies (Shulman, 1981).

Eighteen studies were identified that examined the influence of teachers' mathematical content knowledge on their instructional planning and classroom practice (www.mspkmd.net, 2008). Four of these studies focused on high school teachers. One study found that when teachers with weak content knowledge departed from their instructional materials, they tended to distort the mathematical concepts the students were expected to learn because they chose to increase instruction with inappropriate mathematical representations. Another study found that greater content knowledge strengthened the relationship between positive beliefs about standards-based teaching practices and reported use of these practices.

Mastery of content leads teachers to focus on conceptual teaching more than on procedural teaching (Carpenter, Souder, & Peterson, 1999). Even though the latter is a significant piece in the teaching of mathematics, it has been criticized for being linked to short-term memorization of facts (Carpenter, Fennema, Peterson, Chiang & Loef, 1989).

Conceptual teaching is referred to as "higher order instruction" or "teaching for understanding mathematics" (Carpenter, Fennema and Franke, 1996). Lower achieving students are more likely to experience computational teaching and higher achieving students are more likely to experience conceptual teaching (Clark and Peterson, 1986; Gamoran, 1986; Porter, Kirst, Orthoff, Smithson & Schneider, 1993). Proponents of conceptual-oriented teaching suggest that students do not need to know computational procedures before understanding mathematics (Burrill, 2001).

Hiebert and Stigler (1999) claim that it is difficult for students to understand math once they have learned the rote procedures, and there is better "transfer" when students learn through conceptual understanding rather than memorization. In Japan, students are given time to think about the problem, and the outcome is impressive (Stevenson

and Stigler, 1992). Teachers want their students to be reflective and to gain deeper understanding of mathematics. Each concept and skill is taught with great thoroughness, thereby eliminating the need to teach the concept again later. U.S teachers believe that students learn more effectively if they solve a large number of problems rather than if they concentrate their attention on only a few. "The emphasis is on doing rather than thinking" (Stevenson and Stigler, 1992, p.194)

One common reason offered for the proclivity of U.S teachers' use of more procedural than conceptual teaching is that computational strategies require less in-depth knowledge of mathematics, and teachers in the U.S generally do not have the knowledge and skills required for conceptual teaching on math (Ma & Willms, 1999).

Despite the fact that conceptual teaching is more closely associated with constructivist strategies, it is very much embraced in direct instruction. When the teacher is clear in his explanations and demonstrations, he is attempting to clarify a "concept", and is inviting the students to connect the concept with the algorithm.

Teacher training

Another important factor linked to mastery of content knowledge is teacher education and training. American teachers have master's degrees in teaching methods; Asian teachers hold bachelor degrees in the specific content they teach (Stevenson & Stigler, 1992). Asian teachers' training occurs in their on-the-job experience after graduation from college. In the U.S, training comes to a "near halt after the teachers acquire their teaching certificates" (p.159). "Americans are reluctant to encourage their students to participate at great length during math discussions, because they feel insecure about the depth of their own mathematical training" (p.191).

Darling-Hammond (2000) indicated that the quality of teachers, as measured by whether the teachers were fully certified and had a major in their teaching field, was related to student performance. Measures of teacher preparation and certification were the strongest predictors of student achievement in reading and mathematics -both before and after controlling for student poverty and English language proficiency (Darling-Hammond, 2000).

Clarity and coherence

Teacher clarity has been found to bear a significant, positive relationship to student learning and satisfaction

from the elementary levels through the university level (Metcalf and Cruickshank, 1991; Evertson, Emmer and Brophy, 1991). Two of Rosenshine's and Furst's (1971) eleven characteristics of teacher behavior that showed the strongest relationships with measures of student achievement were clarity of exposition and teacher enthusiasm. Four major themes emerged regarding clarity: (1) the clarity of presentation, (2) the points the teacher makes are easy to understand, (3) the teacher explains concepts clearly and answers questions intelligently, and (4) the lesson is organized. One measure of the clarity of presentation is the amount of time spent answering students' questions, which require an interpretation of what the teacher said. More effective teachers, in terms of student gain in achievement, are able to make the statement once without having to rephrase it only because they did not understand it the first time. Another indicator of clarity is being able to ask students a question once without additional information or more questions interspersed before the students understand and can answer the initial query.

In a study on effective teachers in high poverty schools, Pawlak (2009) found that the most frequently listed effective teacher characteristic was that they

explained well; this was mentioned by over 60% more students than the next most frequently listed characteristics. "Students appreciated explanations that were step-by-step, understandable, repeated in a variety of ways until the students grasped the concept(s), and accompanied by many examples" (p.138).

Teaching cognitive skills

Excellent teachers are concerned with knowing what students understand and how they learn, so they can help students integrate new ideas and transform prior conceptions. Teaching them cognitive strategies will enable them to develop internal procedures that will help them perform higher-level operations (Rosenshine, 1976). Rosenshine (1976) found that processing of new material takes place through a variety of activities such as rehearsal, review, comparing and contrasting, and drawing connections. Such processing strengthens the knowledge network that the student is developing. Asking students to organize information, summarize information, or compare new material with prior material are all activities that require processing and should help students develop and strengthen their cognitive structures.

Marzano et al. (2001) emphasize the importance of teaching students to find similarities and differences in a

specific concept. "Presenting students with explicit guidance in identifying similarities and differences enhances students' understanding of and ability to use knowledge" (Marzano et al., 2001, p.15).

Organizing information

Information can be organized many different ways: by using graphic organizers, or simply by taking notes very efficiently (Marzano, 2003), such as using Cornell Notes. Graphic organizers come in all shapes and forms, and some are even devised by the teachers themselves. Venn diagrams and Frayer models are used extensively in secondary mathematics classes. Although the Frayer model is essentially a vocabulary development tool used for word analysis and vocabulary building, it can be altered to include math problem solving using different representations: numerical, graphic, and verbal. The Frayer model consists of a quadrilateral made up of four quadrants, each one having the following "categories": definition, graphic representation, solution (s) and non-examples. In the middle of the quadrilateral, at the intersection of the quadrants, will be the equation to be solved.

Summarizing and note taking

Marzano (2003) cites four generalizations drawn from the research on note taking that can inform teacher practice: (1) Students should understand that verbatim note taking is probably the least effective way to take notes, since students are so busy taking notes that they don't have time to analyze what they are hearing. (2) Students should regard notes a work-in-progress rather than a final product. Teachers can render note taking more valuable by providing time for students to review and revise their notes, and by helping students identify and correct errors in their notes. (3) Students can use notes as a powerful form of review for tests. (4) The more notes students take, the better. One study showed that there is "a strong relationship between the amount of information taken in notes and students' achievement on examinations" (Marzano et al., 2001, p.43-45).

Multiple representations

Approximately twenty to thirty percent of the school-aged populations remember what is heard; forty percent recall visually the things that are seen or read; the rest rely on manipulatives (Carbo, Dunn & Dunn, 1986). Therefore, an effective instructor owes it to the student,

especially the auditory learner, to explain and clarify a concept as much as possible, and simultaneously support it with relevant examples. The visual learner relies on the information written neatly on the board, and will memorize facts better if they are color-coded. The use of a white board must take precedence over an overhead projector, because the record builds, left to right, as the lesson proceeds, and remains there for the duration of the period, or as long as the students need to absorb the material, view relevant examples and solutions to problems and ask questions. In the U.S classrooms, the overhead projector is preferred because it gives teachers more control over what students are attending to, while in Japan, visual aids provide a cumulative record of the lesson's activities and their results (Stigler & Hiebert, 1999).

Manipulating math sets helps children form important links between real world problems and abstract mathematical notations. The use of algebra tiles to learn factoring of quadratics and completing the square may be useful to kinesthetic as well as visual learners. They are part of reinforcement techniques and active participation of the student, which according to Bloom (1980) accounts for 25 to 40% students' achievement variance.

Concrete-Representational-Abstract techniques (CRA) are the most common example of math instruction incorporating visual representations. It is a 3-part instructional strategy in which the teacher first uses concrete materials (colored chips, base-ten blocks...) to model the mathematical concept, then demonstrates the concept in representational terms (drawing pictures) and finally in abstract or symbolic terms (numbers, math symbols).

The use of manipulatives and other hands-on activities alone does not ensure student understanding of mathematics. Used inappropriately, the use of concrete materials may actually come to replace a student's thinking and interfere with learning (Fennel and Rowan, 2001). The value of using manipulatives, therefore, depends not on whether they are used, but on how they are used with students. An effective teacher mediates students' understanding of the representations and serves as a bridge between the concrete and the abstract.

Explaining proofs

One can use proofs to organize previously disparate results into a unified whole. By organizing a system deductively, one can also uncover arguments that may be fallacious, circular, or incomplete (De Villiers, 1990). By

examining the logical entailments of a concept's definition, one can sometimes develop a conceptual and intuitive understanding of the concept that one is studying. Teaching students how to prove can allow them to independently construct and validate new mathematical knowledge (Yackel and Cobb, 1996).

The *NCTM Standards* argue that by the time the students complete 12th grade, they should recognize proof as fundamental to mathematics, be comfortable with constructing proofs, and be able to determine whether a given argument is a proof. Knuth (2002) interviewed sixteen qualified in-service high school teachers, some with a master's degree, to investigate their conceptions of mathematical proof. When asked about the role of proof, only three teachers indicated that proofs could be used to promote understanding. Knuth concluded that many of these teachers would be unable to effectively meet the NCTM standards.

Teaching Literacy

Teaching literacy in mathematics does not only apply to the early grades. Mathematical terms learned in elementary and middle school are a far cry from those learned in high school. Keeping in mind the diversity that teachers face everyday, accommodating their learning needs

proves to be quite challenging. Words and concepts known as “distractors” in the math teaching community are primarily addressed when teaching students how to read a math question or a word problem. For example, in teaching basic statistics, finding the mean, median, and mode of a group of numbers not only requires applying the rules to come to a solution, but understanding the various meanings of the terms. Here again, finding similarities and differences in meaning proves to be crucial. *Mean* means unkind, signify, and average. *Median* may be confused with the median of a triangle, and *mode* could be construed as fashion to a Spanish speaker. Using Spanish and English cognates may be very helpful in some instances to help students make connections.

Providing guided practice

While not always in agreement about when guidance should be given, both constructivists and proponents of explicit instruction believe that the timing of instructional guidance is important (Schwartz and Branford, 1998). In direct instruction, the best time to provide guidance is as soon as possible—either at the beginning of instruction, or as soon as the learner makes an error. From a constructivist perspective, providing feedback as soon as an error is detected can rob learners of the opportunity to

develop the evaluative skills needed to examine the effects of a problem-solving step, and attempt to repair it in case of error (Mathan and Koedinger, 2003). Large amounts of guidance may produce very good performance during practice, but too much guidance may impair later performance.

In guided practice, activities are initiated under direct teacher supervision. The teacher works the problem step-by-step along with the students. He elicits overt responses from them that demonstrate behavior in objectives. He then slowly releases the students to do more work on their own (they are semi-independent). He then checks for understanding that students were correct at each step. He finally provides specific knowledge of results. This is otherwise known as scaffolding.

Scaffolding

Over the past two decades, an increasing number of educators and researchers have used the concept of scaffolding as a metaphor to explain the role of adults or more knowledgeable peers in guiding children's learning and development (Stone, 1998). The popularity of scaffolding indicates its conceptual significance and practical value for teaching and educational research. Scaffolding should not be seen as only one specific instructional technique. It is a broad term that encompasses many useful and

thoughtful strategies that allows the teacher to break down a task into smaller, more manageable parts in order for the student to understand the full concept (Vygotsky, 1992; Bruner, 1996). If used effectively, over a period of time, scaffolding has the ability to help students cope with the complexity of a task, process how they can accomplish a task, and actually complete the given task, independently. Scaffolding begins at a level that encourages student success and should provide the right amount of support to move students to a higher level of understanding. Scaffolding is used to (1) keep students from straying from the learning objective, (2) organize and support the student's investigations and inquiry, and (3) condition students to accept responsibility for their learning (Bruner, 1976).

Questioning

Effective teachers implement strategies for teaching students how to think, including instruction in study skills, asking higher order questions, and using instructional strategies such as probing, redirection and reinforcement to improve the quality of student responses. Guided practice involves masterful questioning techniques aimed at checking for understanding. Posing the right questions and tweaking their difficulty level to give all

students equal opportunity to answer is no easy task.

Questions have several benefits including:

- ❖ Providing information about prior knowledge and misconceptions.
- ❖ Keeping students' attention on the lesson in progress.
- ❖ Providing an opportunity for review.
- ❖ Providing students the opportunity to monitor their own comprehension and to ask for clarification.
- ❖ Promoting inferences, applications, justifications or solutions to problems.
- ❖ Helping teachers ensure that students are learning the material effectively (Rosenshine, 1976).

An effective math instructor will attempt to address all levels of cognitive thinking in the Bloom hierarchy (Bloom, 1980). From simple knowledge and comprehension to analysis, synthesis and evaluation, questions are varied by type and difficulty level accordingly to assess student mastery of the concept. Here is an example:

1. What is the usefulness of the distributive property?
(Knowledge)
2. Why is the distributive property necessary when dealing with variables? (Analysis)
3. Using algebra tiles and words, construct a problem containing the distributive property. (Application)

4. Write a story problem to match your equation.

(Evaluation)

Lower order questions generally require simple recall or factual answers, whereas higher order questions tend to be more complex and difficult, requiring students to combine facts, form principles, compare, contrast, interpret, and evaluate (Gage, 1976; Rosenshine, 1976). There are obvious qualifications, however. Lower order questions tend to be more effective with younger students who are still acquiring certain cognitive skill processes, with low socio-economic students, and with classes that contain a variety of student abilities (Anderson & Scott, 1978; Gage, 1976). The teacher can use open-ended questions for higher-achieving students (e.g. How should the data be displayed?) and choice questions for lower-achieving students (Should the scores be displayed as a line graph or a bar graph?).

Stevenson and Stigler (1992) maintain that in the States the purpose of asking a question is to get an answer, while in Japan questions are posed to stimulate thought. "Teachers spend a lot of time talking about questions they can pose to the class, which wordings work best to get students involved in thinking and discussing

the material. One good question can keep the whole class going for a long time" (p.195).

Procedural prompts are an excellent questioning tool, and they present the student with an opportunity to assess their own learning (Rosenshine, 1976). Whether the students are studying quadrilaterals or solving cubic functions, their skill at answering these questions is an indicator of their content mastery, or lack thereof.

1. How are rhombi and parallelograms alike?
2. What is the main idea of finding the x-intercepts of the function?
3. What do you think would happen to the graph if the function was quadratic not cubic?
4. In what way is the axis of symmetry related to finding the vertex coordinates?
5. How does moving the parabola two units to the right affect its shape?
6. Compare an isosceles trapezoid and a parallelogram in terms of their consecutive angles.
7. What do you think causes the graph to cross the origin?
8. How does this tie in with what we have learned before?
9. Which one is the best and why?

10. Do you agree or disagree with this statement?

Support your answer.

11. What do you still not understand about the problem?

Assessment

In the current high-stakes educational environment, emphasis is on measurable student learning outcomes. The focus remains on single high-stakes tests, but most assessments of student learning occur in the classroom (Ohlsen, 2007).

Continuous assessment is a key aspect of instructional decision-making. Excellent teachers collect information, interpret those data, and decide what to do next; then they continue to monitor students' progress and adjust the lesson accordingly. In addition to continuous assessment through the teaching-learning process, the student will be assessed at the end of the lesson to determine if the objective has been met. This may be done through traditional assessment approaches (quiz, oral question/answer) or through more authentic approaches (make a poster...) (Gearhart & Saxe, 2004).

Classroom assessment serves many purposes for the teacher: grading, identification of special needs, student motivation, and monitoring instructional effectiveness

(Ohlsen, 2007). Studies by Ohlsen (2007) and Kirtman (2002) found that both beginning and experienced teachers used traditional assessment methods, such as major exams and quizzes, 50% of the time. Major exams are often an assessment tool that teachers use as a cumulative evaluation of student learning at the end of a chapter or unit. Quizzes, on the other hand, serve as an assessment method that allows teachers to assess student learning at a specific point in the learning process (Webb, 2001). Teachers can use the results of tests to determine if remediation or re-teaching is needed for improved student outcomes.

In a quantitative study of 1483 secondary teachers in Virginia, Mc Millan (2001) found that teachers reported high frequencies of use for assessments designed by themselves rather than publisher-created assessments.

In high performing Hispanic schools, many teachers felt that oral assessments removed the pressure from students to perform well on written tests and helped them to: (1) focus more on understanding, (2) develop a mathematics vocabulary, (3) learn how to "think out loud" as they solved problems (Cobb, Wood, & Yackel, 1993, in Pawlak, 2009), and (4) develop a firm foundation of language skills (in both English and Spanish) for later

critical thinking and problem solving use (Reyes et al., 1999, p.101). Other studies, however, have shown that written responses, for assessment purposes, were more representative of students' mastery of content (Reeves, 2004).

Homework

Five studies examined by Marzano on the general effects of homework showed percentile gains of between 1 and 24 (Marzano et al, 2001, p.61). Stronge (2002) found that the quality of the assignments were more important than quantity. Quality assignments provoke thought and allow students to meet the requirements in various creative ways.

Cooper et al. (1989) provide guidelines for homework: (1) Use assignments primarily for instructional and diagnostic purposes, (2) Minimize homework's use for final class grades, (3) Provide information and structure (scaffolding) for students to successfully complete homework without assistance from others, (4) Give a mixture of voluntary and required assignments.

Cooper's (1989) meta-analysis found that for high school students, the positive relation between time on homework and achievement did not appear until at least one hour of homework per week was reported. Then the linear

relation continued to climb to the highest measured interval (more than two hours per night).

Students should receive timely feedback on their independent practice to reinforce their learning and be praised if they have worked well on their own. Effective teachers, as Cooper (1989) suggested, should assign math problems that match students' ability so they can feel successful. Haberman's (1995) star teachers "try to create assignments that youngsters are able to do independently and successfully...Such assignments place the child in the position of expert or explainer to-rather than someone in need of help from-a parent", and "each assignment is special and must pass the same tests of meaningfulness and relevance as in-class activities must" (Haberman, 1995, p.10). Gone should be the days where the teacher announces to the class: "Do problems 1 to 40 on page 55". Math problems are usually numbered by order of difficulty. The first few problems are always simpler to compute than the last few ones. There are also challenge problems towards the end, which should be assigned to the better bunch, if the teacher feels that they are up to the task.

It has become more evident that some teachers often do not require students to think deeply or move beyond the basic knowledge and comprehension levels. The lack of cognitive follow-through in the classrooms leads to superficial thinking, which is ultimately a disservice to students who will be asked to apply their knowledge on a more complex performance oriented task on standardized tests containing open-ended questions. As testing instruments became more sophisticated, short-answer and open-ended, constructed-response items began to appear more frequently on state assessments. Despite the fact that the tests have changed to include a greater emphasis on higher-order thinking with performance-based measures, some teachers have not changed the way they approach their daily instruction (Tankersley, 2000). For this reason, it is in the constructed-response sections where students are having difficulty applying their knowledge. Helping students improve their ability to provide high-quality responses on the constructed-response test items can significantly improve students' scores because each constructed-response item may include many points that could affect the overall scores.

CHAPTER III

METHODOLOGY

Overview

This study primarily looked at correlations among mathematics tests formats. High school students (9th graders, n=394; 10th graders, n= 343) were given the mock CAHSEE in mathematics in two formats: multiple-choice (MC) and constructed response (CR). Each format was made up of the same questions, all of which addressed the California Standards of high school mathematics required to pass the CAHSEE.

Research Questions

The study attempts to explore the relationship between multiple-choice and stem-equivalent constructed response items on the mock CAHSEE in mathematics, and students' scores by gender and language proficiency.

Specifically, in this research, the following questions are being asked:

- 1) What is the relationship between the percentages of students' correct answers on the multiple-choice format and correct answers on the stem-equivalent constructed

responses? What are the differences by gender and language? (Frequency tables were used to calculate proficiency levels).

2) What is the relationship between students' math scores on the multiple-choice standardized mock CAHSEE test and their scores on stem-equivalent constructed responses? (Correlations were run to answer the question).

3) Are there gender differences between the students' scores on the mock CAHSEE multiple-choice questions? Are there gender differences between students' scores on the stem-equivalent constructed responses? (T-tests were performed to answer this question).

4) Are there differences for English Learners between their scores on the multiple-choice questions and their stem-equivalent constructed responses? Are there differences for English Only students between their scores on multiple-choice questions and their stem-equivalent constructed responses? (T-tests were run to answer this question).

5) What is the relationship between the students' mathematics California Standards Test and their scores on the multiple-choice items of the mock CAHSEE? (Correlations were used).

6) What is the relationship between the students' CST scores and their scores on the constructed response tests of the mock CAHSEE? (Correlations were performed).

Data Set

The data set consisted of the ID numbers of 9th and 10th grade students enrolled in Algebra 1, Geometry, and Algebra 2, as well as student demographic information, their scores on both formats of the mock CAHSEE, and their CST math scores from the previous year. The final sample size was 737 after removing those students who moved at the time of testing, absentees, and those in special education.

Student Population Data

The school is the largest of four comprehensive high schools in the Pomona Unified School District. It is located in a predominantly lower middle to lower socio-economic area. One hundred percent of the student body qualifies for free and/or reduced lunch program. Student mobility rate is an ongoing problem. The approximate ethnic make up of the student body is 84% Hispanic, 8% Asian, 3% African-American, and 1% White. There are approximately 800 English Learners. The school is in year 6 of the Program Improvement Placement. It did not meet all of the criteria of the AYP (Adequate Yearly Progress), and its API (Academic Performance Index) in 2010 was 638, compared to

the required score of 800. One subgroup, English Language Learners, did not meet the growth target.

Table #1.

Demographics of Students.

	9 th graders	10 th graders
Hispanic	91.6%	94.2%
Asian	4.4%	2.1%
African American	3.0%	2.9%
White	0.9%	0.8%
Male	48.1%	48.5%
Female	51.9%	51.5%
English Learners	44.4%	43.7%
English Only	55.6%	56.3%
Algebra 1	54.1%	18.1%
Geometry	3.7%	42.9%
Algebra 2	42.2%	39%

The principal provided student ID numbers, information on ethnicity, home/primary language, and student gender. The secretary of the assistant principal in charge of the master schedule gathered CST scores, and enrollment in math classes.

Key Variables

The key variables studied were identified and coded, where necessary, as follows:

1. Demographics: The variable is the primary language: Dichotomous variable of English Learners=1 and English Only=2.
2. Student factors: Variables include:
 - a. Gender: Dichotomous variable of male=1 and female=2.
 - b. Enrollment in math classes: Sub-grouped by Algebra 1, geometry and Algebra 2 and converted to dichotomous variables of yes=1 and no=2.
3. Mathematics scores: Variables include:
 - a. CST scores: Coded as Advanced=5, Proficient=4, Basic=3, Below Basic=2, and Far Below Basic=1.
 - b. Mock CAHSEE multiple-choice (MC) scores: Coded as dichotomous variables of right=1 and wrong=2.
 - c. Mock CAHSEE constructed response (CR) scores: Interpreted as: 1) Raw Score, and 2) Coded as Pass=1 and Fail=2.

Descriptive Statistics

Frequency tables were run to calculate the difference in the percentages of students' scores on the MC and CR items. Those tables also revealed the proficiency levels of

the students: Far Below Basic, Below Basic, Basic, Proficient and Advanced.

Instrumentation

The instrument used is the mock CAHSEE, which is a practice exit examination, given to 9th and 10th graders before the actual CAHSEE. It is developed using state released test items. Each year, students take the practice exit exam and receive a detailed skills analysis two weeks later. Teachers and students use these results to identify areas needing remediation and to provide appropriate instructional and tutoring opportunities.

The CAHSEE. In 1999, the California legislature established the requirement that beginning with the class of 2004, students pass a graduation examination in English Language Arts and Mathematics (SB-2X, written into Chapter 9 of the California Education Code as sections 60850-60859). In July 2003, after the completion of the 2002-2003 CAHSEE testing, the state board of education (SBE) voted to defer the CAHSEE requirement to the class of 2006.

The CAHSEE math covers topics such as statistics and probability, algebra 1, algebra and functions, measurement and geometry, and mathematical reasoning. The standards are at the sixth and seventh grade levels, and cover Algebra 1 as well. The CAHSEE math covers fifty-three academic

content standards: 10 in number sense (Grade 7), 7 in Statistics and Probability (Grades 6 and 7), 10 in Algebra and Functions (Grade 7), 10 in Measurement and Geometry (Grade 7), 6 in Mathematical Reasoning (Grade 7), and 10 in Algebra 1.

Internal Bias and Sensitivity Review. ETS assessment specialists who are specially trained to identify and eliminate questions that contain content or wording that could be construed to be offensive to or biased against members of specific ethnic, racial, or gender groups reviewed every item before it was prepared for content review committees and CDE (ETS, 2008). In addition, the review process promoted a general awareness of and responsiveness to the following:

- 1- Cultural diversity
- 2- Diversity of background, cultural tradition, and viewpoints to be found in the test-taking populations.
- 3- Changing roles and attitudes towards various groups.
- 4- Role of language and setting and changing attitudes toward various groups.
- 5- Contribution of diverse groups to the history and culture of the United States and achievement of individuals within these groups.

Content-related evidence. Content-related evidence refers to the extent to which a student's responses to a given assessment instrument reflects that student's knowledge of the content area that is of interest. For example, an algebra exam should test a student's knowledge using appropriate, relevant math terms, and not complex vocabulary and sentence structures that might unintentionally measure the student's reading comprehension (Moskal, 2000). This would ultimately lead to the teacher misinterpreting the evidence. Content-related evidence is also concerned with the *extent* to which the assessment instrument adequately samples the content domain. A student must be given a problem that would adequately measure his or her *range* of skills.

Construct-related evidence. Reasoning processes are constructs. An isolated correct answer does not provide clear evidence of a student's underlying reasoning process. Since the constructed-response format of any test, notably a mathematics test, provides a clear and precise understanding of a student's reasoning process, it is likely to have a stronger construct-related evidence than a multiple-choice test.

Criterion-Related Evidence. Criterion-related evidence supports the extent to which the students' performance on

the given task may be generalized to other, more relevant activities (Rafilson, 1991).

CAHSEE items were developed to align with the content standards that are representative of the broader content domains: English-language arts, and mathematics. Content validity is determined by a critical review of the items by experts in the field. For the CAHSEE, these reviews are conducted by experts in their designated areas from both the California Department of Education and Educational Testing Service (ETS). For these reviews, ETS senior content staff worked directly with CDE content consultants. The CDE content consultants in the CAHSEE office have extensive assessment experience in their subjects of expertise (California Department of Education, 2008).

After the CAHSEE items were written by ETS-trained item writers, a series of reviews, including reviews by ETS content assessment specialists and external content review committees, were conducted to ensure that each item was measuring the appropriate California content standard and was matched to the item specifications.

The California Standards Tests

Tests are called "standardized" when all students answer the same questions under similar conditions and their responses are scored in the same way. This includes

commercial norm-referenced tests as well as criterion-referenced or standards-based exams. Criterion-referenced tests measure how well a person has learned a specific body of knowledge and skills.

A variation of criterion-referenced testing is "standards based assessment". Many states and districts have adopted content standards (or curriculum frameworks), which describe what students should know and be able to do in different subjects at various grade levels. They also have performance standards that define how much of the content standards students should know to reach the "basic", "proficient", or "advanced" levels in the subject area.

The California Mathematics Standards Tests

Most California Standards Tests reflect the state's academic content standards for the particular grade, with certain exceptions. Mathematics is approached differently. All students in grades 2-6 take the same grade-level test each year. For grades 8-11, the test depends upon the particular math course in which the student is enrolled. The standards assume that 8th graders are registered in Algebra 1, 9th graders in Geometry, and 10th graders in Algebra 2, and these scores are reported. The High School Summative test is only for students who completed that

sequence of courses. Depending on local district curriculum, students in grades 8 through 10 take an alternative test for the first, second, or third year of Integrated Mathematics, an approach that combines algebra, geometry, statistics, and other mathematical knowledge.

The results of the Standards Tests are reported according to the performance level they reach. The California State Board of Education set five benchmarks to indicate a student's proficiency. These levels are Advanced, Proficient, Basic, Below Basic, and Far Below Basic. The percent correct determines the performance level, which differs according to the grade and the level. Since the questions are specifically linked to California's standards, the results have no national comparison (CDE, 2008).

Finding a correlation between the CAHSEE math and the CST math

Cleary, Collins, and Lanier (2008) investigated if a relationship existed between student performance on the California High School Exit Exam (CASHEE) and the California Standards Test. The subjects were all the collective high school sophomores in the state of California from 2005 to 2008. What they found was that, on average, 67% more people passed the CAHSEE than the CST.

Faulk (2008) conducted a study with 1103 student scores from their most recent two years of California Standardized Tests and the California High School Exit Exam (CAHSEE) scores in an effort to identify predictors of success. She found that White and Asian students had the highest passing rates while English Language Learners had the lowest passing rate (25% failed the exam), and both the CST scores for the English Language Arts test and the CST scores for the Mathematics tests predicted the CAHSEE scores.

The Mathematics Standards

All the questions on the tests cover the mathematics standards required to pass the CAHSEE. Eleven questions are related to Number Sense, four are related to Statistics and Probability, four are related to Algebra and Functions, six to Algebra 1, and ten to Measurement and Geometry for a total of 35. The table below lists the content standards tested on the mock CAHSEE with their respective strands and standards sets.

Table #2.

Content Standards of the Mock CAHSEE.

Content Standards	Strand	Standard set
Q1. Scientific notation	Number sense	1.1
Q2. Finding a percentage	Number sense	1.3
Q3. Percent of increase	Number sense	1.6
Q4. Simple interest	Number sense	1.7
Q5. Negative exponents	Number sense	2.1
Q6. Adding fractions and finding common denominators	Number sense	2.2
Q7. Square roots	Number sense	2.4
Q8. Absolute value	Number sense	2.5
Q9. Finding the median	Stat. & Prob.	1.1
Q10. Probability	Stat. & Prob.	3.3
Q11. Probability	Stat. & Prob.	3.3
Q12. Substituting in	Number sense	1.2

rational numbers		
Q13. Interpreting linear graphs	Alg. & Func.	1.5
Q14. Solving square roots with variables	Number sense	2.4
Q15. Interpreting parabolas	Alg. & Func.	3.1
Q16. Solving inequalities	Algebra 1	5.0
Q17. Solving multi-step problems	Algebra 1	5.0
Q18. Finding a relationship between 2 variables	Stat. & Prob.	1.2
Q19. Conversion of units	Meas. & Geom.	1.1
Q20. Scale drawing	Meas. & Geom.	2.3
Q21. Perimeter (inscribed circle)	Meas. & Geom.	2.1
Q22. Area (inscribed circle)	Meas. & Geom.	2.1
Q23. Surface Area	Meas. & Geom.	2.3

Q24. Area of irregular figure	Meas. & Geom.	2.2
Q25. Volume	Meas. & Geom.	2.3
Q26. Area+conversion of units	Meas. & Geom.	2.1
Q27. Pythagorean theorem	Meas. & Geom.	3.3
Q28. Congruence in quadrilaterals	Meas. & Geom.	3.4
Q29. Estimation	Alg. & Func.	2.1
Q30. Finding opposites	Number Sense	2.0
Q31. Absolute value inequality	Algebra 1	3.0
Q32. Distributive property	Algebra 1	4.0
Q33. Interpreting linear graphs	Alg. & Func.	3.3
Q34. System of equations	Algebra 1	9.0
Q35. Multi-step inequality	Algebra 1	5.0

The constructed response questions were devised by copying the multiple-choice questions verbatim and deleting the options. Instructions such as "Explain", "Solve", and "Show work" were added to some of the questions.

1. The radius of the earth's orbit is 150,000,000,000 meters. What is this number in scientific notation?

2. If Freya makes 4 of her 5 free throws in a basketball game, what is her free throw shooting percentage?

3. The cost of an afternoon movie ticket last year was \$4.00. This year, an afternoon movie ticket costs \$5.00. What is the percent increase of the ticket from last year to this year?

4. Sally put \$200.00 in a bank account. Each year, the account earns 8% simple interest. How much interest will be earned in three years?

5. Solve : $(2)^{-4}$

6. Solve: $\frac{5}{6} + \frac{7}{8}$

7. The square root of 150 is between which two numbers?

Show work.

8. If $|x| = 3$, what is the value of x ?

9. From the following numbers, what is the median number?

Explain.

21, 23, 21, 39, 25, 31.

10. To get home from work, Curtis must get on one of the three highways that leave the city. He then has a choice of four different roads that lead to his house. In the diagram below, each letter represents a highway, and each number represents a road.

		Highway		
		A	B	C
Route	1	A1	B1	C1
	2	A2	B2	C2
	3	A3	B3	C3
	4	A4	B4	C4

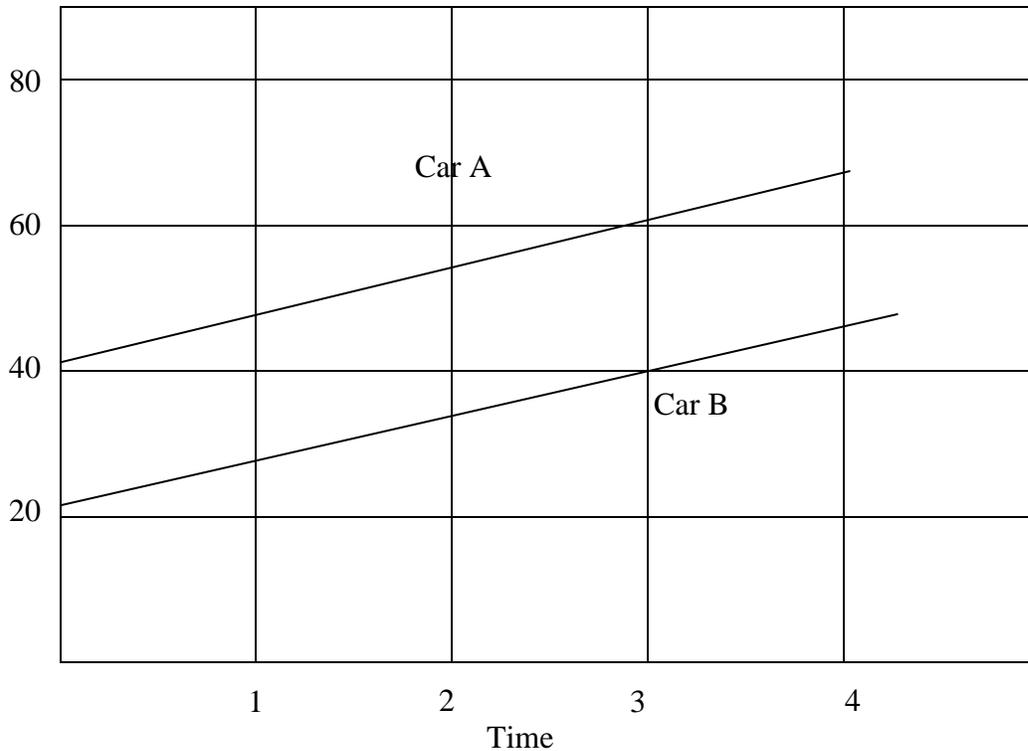
If Curtis randomly chooses a route to travel home, what is the probability that he will travel highway B and route 4?

11. A bucket contains 3 bottles of apple juice, 2 bottles of orange juice, 6 bottles of tomato juice, and 8 bottles of water. If Kira randomly selects a bottle, what is the probability that she will select a drink other than water?

Explain.

12. If $n = 2$ and $x = \frac{1}{2}$, then what is $n(4 - x)$?

13.



After three hours of travel, Car A is about how many kilometers ahead of Car B?

14. Solve: $\sqrt{4x^4} =$

15. Which of the following is the graph of $y = \frac{1}{4}x^2$.

Explain.

4

(Students choose from 4 graphs: One positive parabola, one negative parabola, a linear function, and a cubic function).

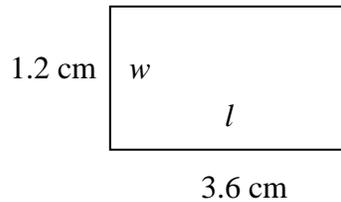
16. In the inequality $2x + \$10,000 \geq \$70,000$, x represents the salary of an employee in a school district. What is the employee's salary? **Use the expressions at least, at most, less than or more than.**

17. Stephanie is reading a 456-page book. During the past 7 days, she has read 168 pages. If she continues reading at the same rate, how many more days will it take her to complete the book?

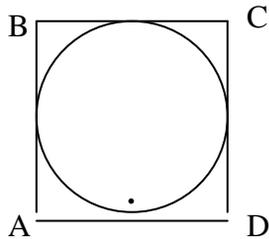
18. Robert's toy car travels at 40 centimeters per second (cm/sec) at high speed and 15 cm/sec at low speed. If the car travels for 15 seconds at high speed and then 30 seconds at low speed, what distance would the car have traveled?

19. A boy is two meters tall. About how tall is the boy in feet (ft) and inches (in)? (1 meter is approximately 39 inches). **Show work.**

20. The actual width (w) of a rectangle is 18 centimeters (cm). Use the scale drawing of the rectangle to find the actual length (l).

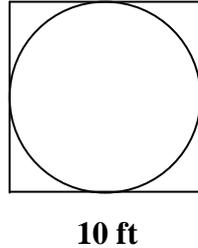


21.



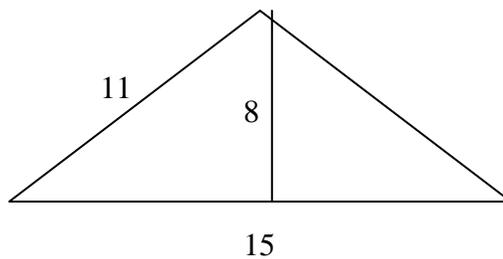
In the figure above, the radius of the inscribed circle is 6 inches (in.). What is the perimeter of square ABCD?

22.



The largest possible circle is to be cut from a 10-foot square board. What will be the approximate area, in square feet, of the remaining board (shaded region)? (The area of a circle is $A = \pi r^2$ and $\pi = 3.14$)

23.

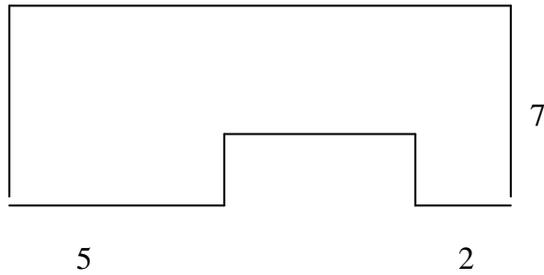


What is the area of the triangle shown above?

24. One-inch cubes are stacked as shown in the drawing below (Figure of a stack of cubes). What is the total surface area?

13

25.

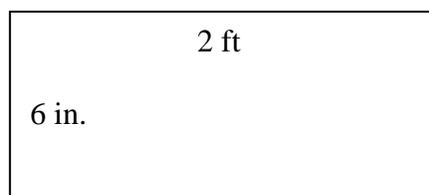


In the figure shown above, all the corners form right angles. What is the area of the figure in square units?

Show work.

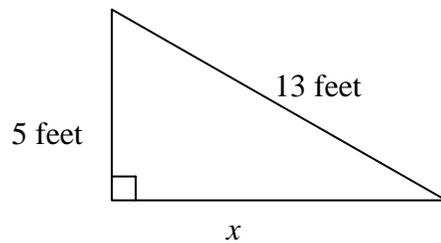
26. The short stairway down below is made of solid concrete (Figure of a stairway). The height and width of each step is 10 inches (in.). The length is 20 inches. What is the volume in cubic inches of the concrete used to create this stairway?

27. The width of the rectangle shown below is 6 inches (in.). The length is 2 feet (ft).

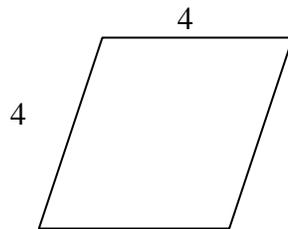


What is the area of the rectangle in square inches?

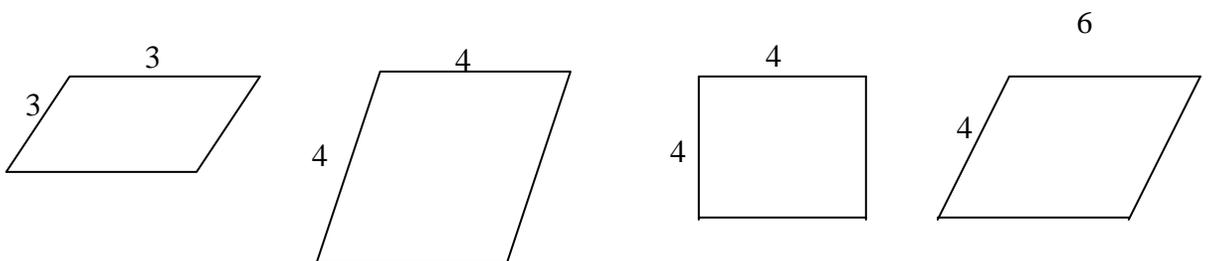
28. What is the value of x in the right triangle shown below? **Show work.**



29.



Which figure is congruent to the figure shown above? **Circle** and explain your choice.



30. The table below shows the number of visitors to a natural history museum during a 4-day period.

Day	Number of Visitors
Friday	597
Saturday	1115
Sunday	1346
Monday	365

Estimate the total number of visitors during this period?

Show your estimate of each number for every day.

Friday =

Saturday =

Sunday =

Monday =

Total =

31. If $x = -7$, then $-x = \dots$

32. If x is an integer, what is the solution to $|x - 3| < 1$?

Show your work.

33. Solve: $4(x + 5) - 3(x + 2) = 14$

34.

$$\begin{cases} 7x + 3y = -8 \\ -4x - y = 6 \end{cases}$$

Solve for x and y . You may use any method (substitution or multiplication).

35. Solve: $9 - 3x > 4(2x - 1)$

Scoring Rubric

The California Mathematics Council rubric is called a general, or holistic, rubric and is used on national or state assessments that must take into account a broad range of mathematical tasks and students. It is aimed at

assigning an overall score rather than a score for particular processes. This type of rubric is appropriate for assessments that are more summative, such as major tests or examinations (Kulm, 1994). "The descriptions of each score are precise enough so that in a short time, teachers can be trained to use the scoring scale with high levels of agreement and reliability" (p.88).

Table #3.

California Mathematics Council Scoring Rubric.

Demonstrated competence
Exemplary response (6 points) - Gives a complete response with a clear, coherent, unambiguous, and elegant explanation; includes a clear and simplified diagram, communicates effectively to the identified audience, shows understanding of the open ended problems' mathematical ideas and processes, identifies all important elements of the problem, may include examples and counterexamples, presents strong, supportive arguments.
Competent response (5 points) - Gives a fairly complete response, fairly clear explanations, includes an appropriate diagram, communicates effectively, shows understanding of the problems' mathematical ideas and

<p>processes, identifies the most important elements of the problem, presents a solid argument.</p>
<p><i>Satisfactory response</i></p>
<p>Minor flaws (4 points)-Satisfactorily completes the problem, a muddled explanation, incomplete argumentation, diagram unclear or inappropriate, understands underlying mathematical ideas, uses mathematical ideas effectively.</p>
<p>Serious flaws (3 points)- Began problem appropriately, failed to complete it, omitted significant parts, failed to show full understanding of mathematical ideas and processes, major computational errors, misuse or lack of use of mathematical terms, used an inappropriate strategy.</p>
<p><i>Inadequate response</i></p>
<p>Begins but fails to complete problem (2 points)- Cannot understand explanation, unclear diagram, shows no understanding of the problem situation, major computational errors.</p>
<p>Unable to begin (1 point)- Inappropriate explanation, diagram misrepresents the problem, copies problem but no attempt at a solution, fails to identify appropriate information.</p>
<p>No attempt (0 points)</p>

Procedures

Thirty-five questions were selected from the mock CAHSEE math booklet (2008 edition) in such a manner that they reflected different standards from every strand. It is customary at this particular school to administer the mock CAHSEE to ninth graders on the day that the tenth graders are taking the actual CAHSEE. The school is on a special schedule because the test is administered all day, from 8 a.m. to 1:30 p.m. Twelve teachers administered the test to 394 Freshmen, who were given the test in constructed response format first, then in multiple-choice format later after a thirty-minute lunch break from 10:30 to 11:00 a.m.

The tenth graders (n=343) were given the test in their math classes two weeks before the CAHSEE. All math teachers agreed to give the multiple-choice format test first on the same day, and waited to give the constructed response test the following week over a period of two days.

All scantrons and constructed response tests had student ID numbers written on them to protect the identity of the students. The students were previously handed a consent form to be signed by their parents, and an assent form to be signed by them agreeing to take the test willingly. They were all aware that it was not just per

school policy that the test was given, but that their scores would be evaluated for the purpose of the study.

Analysis Methods

Correlations were run to explore the relationship between multiple-choice and constructed response scores. Additional correlations were run to examine the relationship between the scores on the CAHSEE and those on the CST mathematics. T-tests were used to investigate the differences in the means of the subgroups on the CAHSEE in both formats. Frequency tables were carried out to examine proficiency levels on each testing format.

CHAPTER IV

RESULTS

The results from the study are presented in the following sections. Correlations, t-tests and descriptive statistics as described in Chapter III are also discussed.

Research Question #1.

What is the relationship between the percentages of students' correct answers on the multiple-choice and their stem-equivalent constructed response items? What are the differences by gender and language?

Table #4.

Pearson Correlation Between Percents of Correct Answers on MC and CR Items.

Correlations

		percentMCcorrect	percentCRcorrect
percentMCcorrect	Pearson Correlation	1	.554**
	Sig. (2-tailed)		.000
	N	742	705
percentCRcorrect	Pearson Correlation	.554**	1
	Sig. (2-tailed)	.000	
	N	705	728

**Correlation is significant at the 0.01 level (2-tailed).

The Sig. 2-tailed level is <.001, which shows that there is a statistical significance between the percentage of correct answers on the MC and CR questions. The relationship is a positive 55.4%, which means that the more likely the student answers correctly on the MC format, the more likely he is to answer correctly on the CR test. Similarly, the higher the likelihood of answering incorrectly on the MC test, the higher the likelihood of answering incorrectly on the CR test. To test the strength of the relationship, the coefficient of determination, which is r^2 is calculated: $r^2 = .31$. It is a moderately strong relationship.

Table #5.

T-Test Results for the Differences between Percents of Students' Correct Answers on MC and CR tests (Gender).

	<u>Boys</u>		<u>Girls</u>		Sig.	t	df	Sig(2-tailed)
	M	SD	M	SD				
MC	47.84	18.45	44.38	15.81	.003	2.74	736	.006
CR	15.88	20.26	12.76	16.22	.000	2.30	723	.02

The t-test revealed a statistically significant difference between the mean percentages of correct answers

on the MC test for boys (M=47.84, s=18.45) and girls (M=44.38, s=15.81), $t(736)=2.74$, $p=.006$, $\alpha=.05$. Since the mean (M) for the boys was greater than the mean (M) for the girls, we can conclude that the percentage of correct answers on the MC test was higher for the boys.

The t-test also revealed a statistically significant difference between the mean percentages of correct answers on the CR test for boys (M=15.88, s=20.26) and girls (M=12.76, s=16.22), $t(723)=2.30$, $p=.02$, $\alpha=.05$. The percentage of correct answers on the CR test was higher for the boys.

Table #6.

T-Test Results for the Differences between Percents of Students' Correct Answers on MC and CR tests (Language).

	EL		EO		Sig.	t	df	Sig.(2-tailed)
	M	SD	M	SD				
MC	38.48	13.71	52.00	17.33	.000	-11.53	736	.000
CR	9.06	11.82	18.48	21.34	.000	-7.13	723	.000

Note. EL= English Learners, EO= English Only students.

The t-test revealed a statistically significant difference between the mean percentages of correct answers on the MC test for ELs (M=38.48, s=13.71) and EOs (M=52.00,

s=17.33), $t(736) = -11.53$, $p < .001$, $\alpha = .05$. Since the mean (M) for the EOs was greater than the mean (M) for the ELs, we can conclude that the percentage of correct answers on the MC test was higher for English Only students.

The t-test also revealed a statistically significant difference between the mean percentages of correct answers on the CR test for ELs (M=9.06, s= 11.82) and EOs (M= 18.48, s=21.34), $t(723) = -7.13$, $p = .000$, $\alpha = .05$. The percentage of correct answers on the CR test was higher for the EOs.

Based on the percents of correct answers, descriptive statistics were also run to compare proficiency levels on multiple-choice and constructed response items for 9th and 10th graders.

Table # 7.

Proficiency Levels of 9th Graders on the MC Test.

	FBB	BB	B	P	A
Boys	36.9%	17.1%	19.3%	11.8%	14.4%
Girls	48.3%	16.9%	21.7%	8.7%	4.3%
EL	62.8%	19.8%	12.8%	3.5%	1.2%
EO	27.5%	14.9%	26.6%	15.3%	15.3%

Table #8.

Proficiency Levels of 9th Graders on CR Test.

	FBB	BB	B	P	A
Boys	66.3%	8.0%	8.6%	8.6%	8.6%
Girls	73.0%	8.3%	8.8%	6.9%	2.9%
EL	87.8%	7.0%	2.9%	1.7%	0.6%
EO	55.7%	9.1%	13.2%	12.3%	9.6%

Table #9.

Comparison of Proficiency Levels of 9th graders on CAHSEE
(in percent).

	FBB		BB		B		P		A	
	MC	CR	MC	CR	MC	CR	MC	CR	MC	CR
Boys	36.9	66.3	17.1	8.0	19.3	8.6	11.8	8.6	14.4	8.6
Girls	48.3	73.0	16.9	8.3	21.7	8.8	8.7	6.9	4.3	2.9
EL	62.8	87.8	19.8	7.0	12.8	2.9	3.5	1.7	1.2	0.6
EO	27.5	55.7	14.9	9.1	26.6	13.2	15.3	12.3	15.3	9.6

Table #10.

Proficiency Levels of 10th Graders on MC Test.

	FBB	BB	B	P	A
Boys	44.5%	15.0%	18.9%	11.0%	11.4%
Girls	46.4%	11.2%	24.6%	11.2%	6.8%
EL	58.4%	18.8%	14.2%	5.8%	2.5%
EO	34.9%	15.9%	16.3%	18.6%	14.2%

Table # 11.

Proficiency Levels of 10th Graders on CR Test.

	FBB	BB	B	P	A
Boys	75.8%	5.1%	9.6%	4.4%	5.0%
Girls	80.1%	5.6%	8.5%	1.8%	4.0%
EL	90.8%	2.6%	4.7%	0.7%	1.4%
EO	67.2%	7.7%	12.8%	5.0%	7.5%

Table # 12.

Comparison of Proficiency Levels of 10th Graders on CAHSEE
(in Percent).

	FBB		BB		B		P		A	
	MC	CR	MC	CR	MC	CR	MC	CR	MC	CR
Boys	44.5	75.8	15.0	5.1	18.9	9.6	11.0	4.4	11.4	5.0
Girls	46.4	80.1	11.2	5.6	24.6	8.5	11.2	1.8	6.8	4.0
EL	58.4	90.8	18.8	2.6	14.2	4.7	5.8	0.7	2.5	1.4
EO	34.9	67.2	15.9	7.7	16.3	12.8	18.6	5.0	14.2	7.5

It is evident that there are significant differences between the scores on both formats for both gender and language. Students tend to perform better on multiple-choice tests than they do on constructed response ones.

Research Question #2.

What is the relationship between students' math scores on the multiple-choice standardized mock CAHSEE test and their scores on stem-equivalent constructed responses?

Table #13.

Pearson Correlation between MC and CR scores.

	CR score	MC score
CR score Pearson r	1	.336**
Sig. (2-tailed)		.000
MC score Pearson r	.336**	1
Sig. (2-tailed)	.000	

Note. **. Correlation is significant at the .01 level.

The Sig. 2-tailed level was $<.001$, which shows that there was a significance between the scores on both formats. The relationship was a positive 33.6%, which means that the higher the student scored on the MC, the more likely he was to score high on the CR test. Similarly, the lower the student scored on the MC test, the more likely he was to score lower on the CR test. The coefficient of determination r^2 is equal to .11. It is a moderate relationship.

Table # 14.

Pearson Correlations between MC and CR questions by Gender and Language.

Boys	.287**
Girls	.401**
EL	.417**
EO	.269**
Total	.336**

Note. EL= English Learners, EO= English Only students.
**. Correlation is significant at the 0.01 level (2-tailed).

There was significance between the 2 variables (MC and CR questions) and the relationship was a positive 28.7% for

the boys, 40.1% for the girls, 41.7% for the English Learners, and 26.9% for English Only students.

Table # 15.

Pearson Correlations between MC and CR questions by Strand.

Number Sense	.765**
Statistics & Probability	.578**
Algebra 1	.276**
Algebra&Functions	.525**
Measurement and Geometry	.545**

Note. **. Correlation is significant at the .01 level.

There was a significant positive relationship between MC and CR questions.

Additional statistics were run to find correlations between MC and CR scores on each question of every strand. These tables can be found in Appendix A.

Research Question #3.

What are the gender differences between the students' scores on multiple-choice and stem-equivalent constructed response questions?

351 boys and 386 girls took the test. A t-test was run to determine the significant differences between the means of the boys and the girls.

Table # 16.

T-Test Results for Relationships between MC and CR scores by Gender.

	<u>Boys</u>		<u>Girls</u>		Sig.	t	df	Sig.(2-tailed)
	M	SD	M	SD				
MC	2.42	1.43	2.15	1.22	.000	2.75	735	.006
CR	32.94	45.08	32.76	43.56	.721	.052	719	.960

The t-test revealed a statistically significant difference between the means of MC scores for boys (M=2.42, s=1.43) and girls (M=2.15, s=1.22), $t(735)=2.75$, $p=.006$, $\alpha=.05$. Since the mean (M) for the boys was greater than the mean (M) for the girls, we can conclude that the scores on the MC test were higher for the boys.

The t-test failed to reveal a statistically significant difference between the means of CR scores for boys (M=32.94, s=45.08) and girls (M=32.76, s=43.56), $t(719)=.052$, $p=.960$, $\alpha=.05$. The significance was .960, which is greater than .05. We can assume that variances were approximately equal.

Research Question #4.

What are the differences for English Learners and English Only students between their scores on the multiple-choice questions and their stem-equivalent constructed responses?

An independent t-test was run to investigate differences between the means of English Language learners (N= 326) and English Only students (N= 402).

Table # 17.

T-Test Results for Relationships between MC and CR scores by Language.

	<u>EL</u>		<u>EO</u>					
	M	SD	M	SD	Sig.	t	df	Sig.(2-tailed)
MC	1.71	.99	2.73	1.39	.000	-11.28	735	.000
CR	26.31	35.45	38.21	49.76	.000	-3.62	719	.000

Note. EL= English Learners, EO= English Only students.

The t-test revealed a statistically significant difference between the means of MC scores for EL (M=1.71, s=.99) and EO (M=2.73, s=1.39), $t(735) = -11.28$, $p < .001$, $\alpha = .05$. The scores on the MC test were higher for the English Only students.

The t-test also revealed a statistically significant difference between the means of CR scores for EL (M=26.31, s=35.45) and EO (M=38.21, s=49.76), $t(719) = -3.62$, $p < .001$,

$\alpha=.05$. Since the mean (M) for the EO was greater than the mean (M) for the EL, we can conclude that the scores on the CR test were higher for the English Only students.

Research Questions #5 and #6.

What is the relationship between the students' mathematics California Standards Test scores and their scores on the multiple-choice and constructed response items on the mock CAHSEE?

Table # 18.

Pearson Correlation between CAHSEE and CST scores.

	CR score	MC score	CST score
CR score r	1	.336**	-.036
Sig. 2-tailed		.000	.353
N	725	702	682
MC score r	.336**	1	.524**
Sig. 2-tailed	.000		.000
N	702	741	698
CST score r	-.036	.524**	1
Sig. 2-tailed	.353	.000	
N	682	698	860

The p value for the MC/CST scores was $<.001$, which shows significance between the MC score and the CST score. The relationship was a positive 52.4%. The p value for the CR/CST scores was .353, which is greater than .05. There was no significant correlation found between constructed response scores and CST scores ($r=-.036$).

Table # 19.

Pearson Correlations between CST and CAHSEE scores for Gender and Language.

	MC/CST	CR/CST
Boys	.570** (r^2 .32)	-.008
Girls	.459** (r^2 .21)	-.068
EL	.371** (r^2 .14)	.095
EO	.524** (r^2 .27)	-.146**
Total	.524** (r^2 .27)	-.036

Note. EL= English Learners, EO= English Only students.
 **. Correlation is significant at the 0.01 level (2-tailed).

There was a significant positive relationship between MC scores on the CAHSEE and CST math scores. The coefficient of determination r^2 shows a moderate to moderately strong relationship between both MC and CST scores. There was, however, a significant negative

correlation between the CR scores and CST scores for English Only students, which means that the higher they tended to score on the CST test, the lower their score on the CR, and vice versa. There were no significant correlations between the CR scores and the CST math scores for the rest of the independent variables.

The implications of all the results presented above are discussed in Chapter Five. Limitations of the study are mentioned as well, and recommendations for future math instructors are also suggested.

CHAPTER V

CONCLUSION

The purpose of the study was to explore relationships between students' scores on multiple-choice and stem-equivalent constructed response questions on the mock CAHSEE in mathematics in a low performing, predominantly Latino high school. The students' scores on the California Standards Test in mathematics were also correlated with their scores on the mock CAHSEE. Frequency tables were run to investigate percentages of students scoring at various levels of proficiency on both formats. Empirical data were disaggregated and analyzed by gender and by language (English Learners versus English Only). Statistical analyses were performed using correlations, T-tests, and descriptive statistics.

The sample consisted of 737 students enrolled as freshmen and sophomores in algebra 1, algebra 2 and geometry. The majority of the students were Latinos, but there were also Asian students of different ethnic backgrounds, African American students, and some white students. Due to the insignificant percentage of non-Latinos (9%), the ethnicity variable, which was initially considered in the study, had to be discarded.

The test consisted of eleven questions related to Number Sense (NS), four in Statistics and Probability (S&P), four in Algebra and Functions (A&F), six in Algebra 1 (Alg1), and ten in Measurement and Geometry (MG). The California Mathematics Council rubric was used to score the constructed response questions.

Results and implications of the study will be discussed in this chapter.

Research Findings

Research Question #1.

What is the relationship between the percentages of students' correct answers on the multiple-choice and their stem-equivalent constructed response items? What are the differences by gender and language?

The correlation was a positive .554 at the .01 level, and the coefficient of determination r^2 was .31, which indicates a moderately strong relationship.

A t-test revealed a statistically significant difference between the mean percentages of correct answers on the MC test for boys and girls. The mean for the boys was greater than the mean for the girls, so the percentage of correct answers on the MC test was higher for the boys.

The t-test also revealed a statistically significant difference between the mean percentages of correct answers on the CR test for boys and girls. The percentage of correct answers on the CR test was higher for the boys.

Another t-test revealed a statistically significant difference between the mean percentages of correct answers on the MC test for English Learners and English Only students. The mean for the EOs was greater than the mean for the ELs, so the percentage of correct answers on the MC test was higher for English Only students.

The t-test also revealed a statistically significant difference between the mean percentages of correct answers on the CR test for ELs and EOs. The percentage of correct answers on the CR test was higher for the EOs.

A look at the proficiency levels revealed significant differences between the percentages on both formats for both gender and language. Even though a moderately strong relationship was found between the percentages in both formats, the data suggest that students tend to perform better on multiple-choice tests than they do on constructed response ones.

Research Question #2.

What is the relationship between students' math scores on the multiple-choice standardized mock CAHSEE test and their scores on stem-equivalent constructed responses?

Statistically significant positive correlations were found between the multiple-choice and the constructed response total scores ($r=.336^{**}$). The coefficient of determination r^2 was equal to .11, which indicates a moderate relationship.

Correlations were also run to examine the relationship between MC and CR items on every strand of mathematics. Number Sense showed the most significant positive correlation ($.765^{**}$), followed by Statistics and Probability ($.578^{**}$), then Measurement and geometry ($.545^{**}$), Algebra and Functions ($.525^{**}$), and finally Algebra 1 ($.276^{**}$).

Additional correlations were run for every question on every strand. All number sense questions showed a significant relationship between both formats, except question number 1 (scientific notation), which found no correlation for the English Learners. All questions related to Statistics and probability showed significant correlations on both formats for all independent variables.

In the Algebra 1 strand, results were mixed. No significant correlations were displayed for English Learners on questions 30 (estimation), and there were no significant correlations for question 35 (inequality) for boys. No correlations were found on question 32 (absolute value inequality) for all independent variables.

In Measurement and Geometry, most questions displayed significant positive correlations, except for English Learners whose scores revealed no relationships for questions 20 (scale drawing), 22 (area problem), 24 (surface area), and 28 (Pythagorean Theorem).

Algebra and Functions items showed significant positive correlations for all independent variables.

Research Question #3.

What are the gender differences between the students' scores on multiple-choice and stem-equivalent constructed response questions?

The t-test revealed a statistically significant difference between the means of MC scores for boys and girls. The scores on the MC test were higher for the boys.

The t-test failed to reveal a statistically significant difference between the means of CR scores for boys and girls. The significance was .960, which is greater

than .05. We can assume that variances were approximately equal.

Research Question #4.

What are the differences for English Learners and English Only students between their scores on the multiple-choice questions and their stem-equivalent constructed responses?

The t-test revealed a statistically significant difference between the means of MC scores for English Learners and English Only students. The scores on the MC test were higher for the English Only students.

The t-test also revealed a statistically significant difference between the means of CR scores for EL and EO. Since the mean for the EO was greater than the mean for the EL, we can conclude that the scores on the CR test were higher for the English Only students.

Research Questions #5 and #6.

What is the relationship between the students' mathematics California Standards Test scores and their scores on the multiple-choice and constructed response items on the mock CAHSEE?

The p value for the MC/CST scores was $<.001$, which shows significance between the MC score and the CST score. The relationship was a positive 52.4%. The p value for the CR/CST scores was .353, which is greater than .05. There

was no significant correlation found between constructed response scores and CST scores ($r=-.036$).

Limitations Of The Study

One major challenge at the onset of the study was to have consistency in the administration of the tests. There were two sets of teachers: a) those who proctored the mock CAHSEE for the 9th graders in one day, with a half hour break between giving the test in CR format first, then in MC format, b) the 10th grade teachers who volunteered to give the tests to their students in MC format on a given day, then in CR format the following week.

The 9th graders were more controlled due to the fact that they were required to attend on the day that the 10th graders were taking the actual CAHSEE in math. Attendance was very good, and the proctors had to monitor them following state guidelines, so cheating was minimized, and the classroom environment was restrained.

The 10th graders took the mock CAHSEE in their respective math classes on a regular day, ten days before they were to take the actual CAHSEE. There were many students who were absent on the days they had to take both tests. Some took one test but failed to take the other. It is uncertain how teachers monitored the students, since the person who conducted the study was not present at the time

of the testing. One teacher failed to turn in all of his tests. Two teachers turned in a few incomplete MC tests, which resulted in missing data, and skewed scores.

There were few CST scores (about 2% of the total scores) that were not available for some students. It was not known whether the student had taken the test but scores were never reported, or if the student had never taken the test.

It was originally the intent of the researcher to examine the ethnicity variable but the number of Asian students, African American students and White students was significantly negligent compared to the Latino students, so the ethnicity variable was dropped.

Implications

It is important to investigate the extent of proficiency students have in reading the math questions, solving the problems, and writing about their thinking processes. Differences were evident in the proficiency levels which were gleaned from the percentages of correct answers on both testing formats: on the constructed response items, more students scored at the far below basic level and less students scored at the proficient and advanced levels, while there seemed to be more success on

the multiple-choice questions. On the constructed response test, there were many questions left blank. The percentage of students scoring a 0 on every question of every strand was tabulated. The average percentage of 0 scores was then calculated for every strand: 15.26% for Number Sense questions, 16% for Statistics and Probability, 35.3% for Algebra 1 items, 18.5% for Algebra and Functions, and 32.4% for Measurement and Geometry. This should alert teachers that students, especially those enrolled in low performing schools, and who are English Learners, need to be given performance tasks, and be encouraged to write their thinking processes in order for their skills to be more properly assessed.

Marzano et al. (2001) stress that explaining their thinking helps students to enhance their understanding of the experimental inquiry process and their use of the steps involved. Also, the *range* of cognitions - such as knowledge, procedures, images and skills - that can be elicited by CR items is greater than the range of MC items (Martinez, 1999).

Traub and McRury (1990) reported that students had more positive attitudes towards multiple choice tests in comparison to free response tests because they thought that these tests were easier to prepare for, easier to take, and

thus will bring in relatively higher scores. It is the researcher's belief that since teachers are being held accountable for their teaching by virtue of their test scores, they may prefer to give the students tests on which they are more likely to be successful. This is where they should step up their teaching practices and empower students by training them to become capable critical thinkers, and motivating them to participate in hands-on problem solving activities.

Birenbaum and Feldman (1998) discovered that students with a deep study approach tended to prefer essay type questions, while students with a surface study approach tended to prefer multiple-choice formats. As a result of the research findings, it behooves the teachers to initiate changes in students' study habits, notably English Language Learners, and encourage them to favor open-ended formats, while providing language accommodations. English Learners have literacy challenges when processing their strategies, and some critics suggest that, for ELLs, the fairest approach is to focus almost exclusively on the reduction of language in the text (Abedi, 2008).

Hiebert and Stigler (1999) claim that it is difficult for students to understand math once they have learned the rote procedures, and there is better "transfer"

when students learn through conceptual understanding rather than memorization. In Japan, students are given time to think about the problem, and the outcome is impressive (Stevenson and Stigler, 1992). U.S teachers believe that students learn more effectively if they solve a large number of problems rather than if they concentrate their attention on only a few. "The emphasis is on doing rather than thinking" (Stevenson and Stigler, 1992, p.194).

Students who are given the opportunity to show and explain their mathematical reasoning have a better chance of earning points on a well thought out process, even if the ultimate response was wrong due to an arithmetic error. It would be evident to the teacher that the student knew how to work out the problem, but had the misfortune of placing a negative sign where a positive sign was due. Such an error would not be obvious on a multiple-choice test, which only displays the wrong answer, and does not reveal how the mistake came about. According to the NCTM (1991), although the commonly used MC format may yield important data, it can have a negative impact on how students are taught and evaluated at the school level because: a) Student scores are generated solely on the basis of right and wrong answers with no consideration or credit given to students' strategies, b) Routine timing measures how

quickly students can respond but not necessarily how well they think, and c) Mathematics tools such as calculators and measurement devices are not permitted (1991, p.8).

Willingham and Cole (1997) reviewed national and state assessment results and concluded that writing often appeared to play a role in gender format score differences. The research they reviewed suggested writing skills and fluency differences as possible factors in the female advantage on CR tasks. They also reported that requested discussion and explanation of responses consistently showed female advantages. In this study, it was revealed that girls left as many blank answers as the boys and earned an approximate equal amount of low scores on the constructed responses.

Recommendations

Integrating open-ended math problems, as well as implementing performance tasks, which promote cognitive thinking, will prepare the students to be more confident and efficient problem solvers. Teachers must strive to incorporate multiple choice and constructed response items on their tests to assess skills as well as literacy. Douglas Reeves, chairman and founder of the Center for Performance Assessment and the International Center for

Educational Accountability, has said that "even if the state test is dominated by lower-level thinking skills and questions are posed in a multiple-choice format, the best preparation for such tests is not mindless testing drills, but extensive student writing, accompanied by thinking, analysis, and reasoning" (2004, p. 92). It is crucial that teachers give *all* students equal opportunities to prove their potential, and dispel misconceptions that low ability students can only handle MC questions, while high ability students can take on answering open-ended questions, as Fleming (1998) found in her study.

Development of skills required for academic achievement can be influenced by instructional design. By understanding and incorporating open-ended activities into the regular instructional program, teachers can feel confident that their students will quickly become better prepared for meeting the challenges they will face on the constructed-response sections of assessments.

We need teachers who can *teach* the content, not just *know* the content. Teachers must implement literacy skills and academic discourse in their classes so students can express what they know and write it clearly and persuasively. Teachers must incorporate open-ended

activities, and assign performance tasks into their regular instructional program.

If we want our students to be proficient writers in mathematics, we must give them the opportunity to write and express their ideas and their reasoning. Students are better prepared to take standardized multiple choice tests if they are trained to be test-wise AND given the opportunity to answer open-ended questions. We will have students who are strategic learners as well as capable problem solvers.

APPENDIX A

Table 20.

Pearson Correlations between MC and CR Questions for Number Sense.

Question #	Boys	Girls	EL	EO
1	.470**	.142**	.079	.530**
2	.591**	.594**	.575**	.611**
3	.561**	.418**	.394**	.563**
4	.475**	.295**	.332**	.427**
5	.471**	.487**	.433**	.480**
6	.687**	.620**	.545**	.676**
7	.486**	.545**	.441**	.564**
8	.423**	.354**	.334**	.389**
12	.379**	.335**	.337**	.335**
14	.386**	.362**	.360**	.342**
17	.480**	.465**	.321**	.570**

Note. EL= English Learners, EO= English Only students.

** . Correlation is significant at the 0.01 level (2-tailed).

* . Correlation is significant at 0.05 level (2-tailed).

Table 21.

Pearson Correlations between MC and CR Questions for Statistics and Probability.

Question #	Boys	Girls	EL	EO
9	.600**	.694**	.591**	.673**

10	.500**	.385**	.445**	.388**
11	.404**	.378**	.386**	.330**
18	.147**	.380**	.035	.473**

Note. EL= English Learners, EO- English Only students.
 **. Correlation is significant at the 0.01 level (2-tailed).
 *. Correlation is significant at the 0.05 level (2-tailed).

Table #22.

Pearson Correlations between MC and CR Questions for Algebra 1.

Question #	Boys	Girls	EL	EO
16	.306**	.296**	.208**	.346**
30	.241**	.153**	.088	.210**
31	.352**	.283**	.254**	.355**
32	.098	-.096	.019	-.022
34	.204**	.171**	.123*	.202*
35	.091	.186**	.163**	.115*

Note. EL= English Learners, EO= English Only students.
 **. Correlation is significant at the 0.01 level (2-tailed).
 *. Correlation is significant at the 0.05 level (2-tailed).

Table 23.

Pearson Correlations between MC and CR Questions on Measurement and Geometry.

Question #	Boys	Girls	EL	EO
19	.351**	.444**	.317**	.417**
20	.467**	.199**	.102	.576**
21	.504**	.448**	.344**	.525**

22	.236**	-0.40	-.082	.272**
23	.438**	.312**	.219**	.454**
24	.359**	.113*	.033	.358**
25	.140	.292**	.227**	.146*
26	.156**	.048	.193**	.107*
27	.367**	.388**	.156**	.465**
28	.166**	.221**	.100	.368**

Note. EL= English Learners, EO= English Only students.

** . Correlation is significant at the 0.01 level.

* . Correlation is significant at the 0.05 level.

Table 24.

Pearson Correlations between MC and CR Questions on Algebra and Functions.

Question #	Boys	Girls	EL	EO
13	.588**	.456**	.534**	.457**
15	.537**	.622**	.539**	.610**
29	.362**	.280**	.251**	.375**
33	.378**	.186**	.198**	.304**

Note. EL= English Learners, EO= English Only students.

** . Correlation is significant at the 0.01 level (2-tailed).

* . Correlation is significant at 0.05 level (2-tailed).

Bibliography

- Abedi, J. (2002). Standardized Achievement Tests and English Language Learners: Psychometrics Issues. *Educational Assessment, 8*(3), 231-257.
- Abedi, J. & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*, 219-234.
- Ackerman, T., & Smith, P. (1998). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement, 12*(2), 117-128.
- Aiken, Lewis R. (1972). Language factors in learning mathematics. *Review of Educational Research, 42*(3), 359-385.
- Becker, W.E., & Johnston, C. (1999). The relationship between multiple-choice and essay response questions in assessing economics understanding. *Economic Record, 75* (231), 348-357.
- Ben-Chaim, D., & Zoller, U. (1997). Examination-type preferences of secondary school students and their teachers in the science disciplines. *Instructional Science, 25*(5), 347-367.

- Bennett, R.E., Ward, W.C., Rock, D.A, & Lahart, C. (1990). *Toward a Framework for Constructed-Response Items*. Princeton, New Jersey. Education testing Service
- Berliner, D. (1986). Should students be made test-wise? *Instructor*, 95(6), 22-23.
- Birenbaum, M., & Tatsuoka, K.K. (1987). Open-ended versus multiple-choice response formats- It does make a difference for diagnostic purposes. *Applied Psychological Measurement*, 11, 385-395.
- Bohlin, C.F. (1994). Learning style factors and mathematics performance: Sex related differences. *International Journal of Educational Research*, 21, 387-397.
- Bracey, G. W. (2002). *Put to the test: An educator's and consumer's guide to standardized testing*. Bloomington IN: Phi Delta Kappa International.
- Brahier, D.H (2001). *Assessment in middle and High School Mathematics*. Archmont, N.Y: Author
- Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement*, 29(3), 253-271.
- Bruner, J. (1996). *The culture of education*. Harvard University Press.
- Burrill, G. (2001). Mathematics Education: The Future and The Past Create A Context for Today's Issues. In Loveless,

T. (Eds). The Great Curriculum Debate: How Should We Teach Reading and Math, 25-41.

Burton, N.W. (1996). How have changes in the SAT affected women's math scores? *Educational Researcher*, 15(4), 5-9.

Carbo, M., Dunn, R. & Dunn, K. (1986). *Teaching students through their individual learning styles*. Englewood Cliffs, NJ: Reston Book; Prentice Hall.

Carnine, D. & Gersten, R. (1982). Effective mathematics instruction for low-income students: Results of longitudinal field research in 12 school districts. *Journal for Research in Mathematics Education*, 13(2), 145-152.

Carpenter, P., Fennema, E., Peterson, P. L., Chiang, C. P., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal*, 26(4), 499-531.

Carpenter, P., Fennema, E., & Franke, M.L. (1996). Cognitively Guided Instruction: A Knowledge Base for Reform in Primary Mathematics Instruction. *Elementary School Journal*, 97, 3-20.

Cizek, G. J. (1998). *Filling in the blanks: Putting standardized tests to the test*. Washington D.C.: The Thomas B. Fordham Foundation.

Cizek, G. J. (2001, Winter). More unintended consequences of high-stakes testing. *Educational Measurement, Issues and Practice*, 20(4), 19-28.

Cleary, B., Collins, D.J., Lanier, B. (2008). What is the relationship between student performance on the California High School Exit Exam (CAHSEE) and the California Standards Tests. Retrieved from the World Wide Web on December 1, 2010, <http://www.google.com>.

Coltrane, B. (1992). English Language Learners and High Stakes Tests: An overview of the Issues. ERIC Digest. Available: Doc. EDO-FL-02-07.

Cooper, H., Robinson, J., & Patall, E. (1989). Does Homework Improve Academic Achievement? A Synthesis of Research, 1987-2003. *Review of Educational Research*, 76(1), 1-62.

Crocker, L. & Schmitt, A. (1987). Improving multiple choice test performance for examinees with different levels of test anxiety. *The Journal of Experimental Education*, 55(4), 201-205.

Cronbach, L.J. (1988). *Five perspectives on validity argument*. In H. Wainer and H.I. Braun (Eds). Lawrence Erlbaum, Hillsdale, NJ.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Educational Policy Analysis, 8*, 1-48.

Delgado, A.R., & Prieto, G. (2003). The effect of item feedback on multiple choice test responses. *British Journal of Psychology, 94*(1), 73-85.

DeMars, C. E. (1998). Gender differences in mathematics and science on a high school proficiency exam: The role of response format. *Applied Measurement in Education, 11*(3), 279-299.

De Villiers, M.D. (1990). The role and function of proofs in mathematics. *Pythagoras, 24*, 17-24.

Dickinson, A. C., Friedman, M. I., Hatch, C. W., Jacobs, J. E., Nickerson, A. B., & Schnepel, K. C. (2002). *Educators' Handbook on Effective Testing*. Columbia, SC: Institute for Evidence-Based Decision-Making in Education.

Dolly, J.P., & Williams, K.S. (1986). Using test-taking strategies to maximize multiple-choice test scores. *Educational and Psychological measurement, 46*, 619-625.

Dolly, J.P., & Williams, K.S. (1983). Teaching testwiseness. Eric Digest . Available: Doc. ED 241562.

Doolittle, A.E., & Cleary, T.A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement, 24*, 157-166.

Dufresne, R.J., Leonard, W.J., & Gerace, W.J (2002). Making sense of students' answers to multiple-choice questions.

The Physics Teacher, 40, 174-180.

Educational Testing Service (2009). *Guidelines for the Assessment of English Language Learners*. Retrieved from the World Wide Web on November 30th, 2010, <http://www.ets.org>.

Edwards, B.D. (2007). An examination of factors contributing to a reduction in subgroup differences on a constructed-response paper-and-pencil test of scholastic achievement. *Journal of Applied Psychology*, 92(3), 794-801.

Evertson, C., Emmer, E., & Brophy, J. (1980). Predictors of Effective Teaching in Junior High Mathematics Classrooms. *Journal for Research in Mathematics Education*, 11(3), 167-178.

Fagley, N.S. (1987). Positional response bias: Its relationship to testwiseness and guessing strategy. *Journal of Educational Psychology*, 79, 95-97.

Fagley, N.S., Miller, P.M., & Downing, R. (1990). *Convergent and discriminant validity of the experimental test of wiseness*. Paper presented at the annual meeting of the American Psychological Association, Boston. Retrieved from the World Wide Web on October 2nd, 2010, <http://www.google.com>.

Faulk, L. (2008). Predicting mathematical proficiency : An Examination of Standardized Test Scores, Gender, and Ethnicity in Order to Predict CAHSEE Passing Rates. Retrieved from the World Wide Web on December 2, 2010: <http://www.google.com>.

Fennell, F., & Rowan, T. (2001). Representation: An important process for teaching and learning mathematics. *Teaching Children Mathematics*, 7(5).

Fleming, K., Ross, M., Tollefson, N., & Green S.B. (1998). Teachers' choices of test item formats for classes with diverse achievement levels. *The Journal of Educational Research*, 91(4), 222-228.

Frary, R.B. (1995). "The none of the above" option: An empirical study. *Applied Measurement in Education*. 4(2), 115-124.

Frary, R.B., Cross, L.H., & Lowry, S.R (1977). Random guessing, correction for guessing, and reliability of multiple-choice test scores. *The Journal of Experimental Education*, 46(1), 11-15.

Fredricksen, N. (1984). The real test bias. *American Psychologist*, 39, 193-202.

Frey, B.B., Petersen, S.E., Edwards, L.M., Pedrotti, J.T. & Peyton, V. (2003). Toward a consensus list of item-writing

rules. Presented at the Annual Meeting of the American Educational Research Association, Chicago.

Gallagher, A.M. (1992). *Sex differences in problem solving strategies used by high scoring examinees on the SAT-M*. New York: College Entrance Examination Board.

Gamoran, A., Porter, A., Smithson, J., & White, P. (1997). Upgrading High School Math Instruction: Improving Learning Opportunities For Low-Achieving, Low-Income Youth.

Educational Evaluation and Policy Analysis, (19), 325-338.

Garner, M. & Engelhardt, J.G. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education*, 12(1), 29-51.

Gay, L.R. (1980). The comparative effects of multiple-choice versus short-answer tests on retention. *Journal of Educational Measurement*, 17(1), 45-50.

Gearhart, M., & Saxe, G. B. (2004). When Teachers Know What Students Know: Integrating Mathematics Assessment. *Theory Into Practice*, 43(4), 304-313.

Geiger, M.A (1997). An examination of the relationship between answer changing, testwiseness, and examination performance. *The Journal of Experimental Education*, 66(1), 49-60.

Gellman, E., & Berkowitz, M. (1993). Test-item type: What students prefer and why. *College Student Journal*, 27(1), 17-26.

Gutierrez, R. (2002). Beyond essentialism: The complexity of language in teaching mathematics to Latino/a students. *American Educational research Journal*, 39(4), 1047-1088.

Haberman, M. (1995). *Star Teachers of Children in Poverty*. Indianapolis: Kappa Delta Pi.

Haldyna, T. M. (1997). *Writing Test Items to Evaluate Higher Order Thinking Skills*. Allyn Bacon, Boston, MA.

Haladyna, T.M., & Downing, S.M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1), 37-50.

Hambleton, R.K., & Murphy, E. (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education*, 5(1), 1-16.

Hamilton, L. (1994). An Investigation of Students' Affective responses to Alternative Assessments Formats.

Paper presented at the Annual Meeting of the National Council on Measurement in Education. New Orleans, LA.

Retrieved from the World Wide Web on October 3rd, 2010, <http://www.altavista.com>.

Hancock, G.R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response

test formats. *Journal of Experimental Education*, 62(2), 143-157.

Heubert, J. P. (2009). High-Stakes Testing: Opportunities and Risks for Students of Color, English-Language Learners, and Students with Disabilities. Retrieved from the World Wide Web on November 27th, 2010, <http://www.google.com>.

Hiebert, J. & Stigler, J. & (1999). *The Teaching Gap*. The Free Press, New York, NY.

Hirsch, E.D. (2007). *The Knowledge Deficit*. Houghton Mifflin Company, New York.

Holder, W.W., & Mills, C.N. (2001). Pencils down, computer up: The new CPA exam. *Journal of Accountancy*, 191(3), 57-60.

Hollingworth, L., Beard, J. J., & Proctor, T.P (2007) An investigation of item type in a standards-based assessment. *Practical Assessment Research and Evaluation*, 12(18), 1-13.

Hunter, M. (1979). Diagnostic teaching. *The Elementary School Journal*. 80(1), 41-46.

Kimball, M. (1989). A new perspective on women's math achievement. *Psychological Bulletin*, 105, 198-214.

Kirtman, L. (2002). *Policy and Practice: Restructuring Teacher's Work*. Retrieved from the World Wide web on February 28th, 2010, <http://epaa.asu.edu/epaa>.

Knapp, M.S., Adelman, N.E., Marder, C.C, McCollum, H.,
Needles, M.C., Padilla, C. et al. (1995). *Teaching for
meaning in high-poverty classrooms*. New York, NY: Teacher
College Press.

Knowles, S.L. & Welch, C.A. (1992). A meta-analytic review
of item discrimination and difficulty in multiple-choice
items using "none of the above". *Educational and
Psychological Measurement*, 52, 571-577.

Knuth, E. (2002). Secondary school mathematics teachers'
conceptions of proof. *Journal for Research in Mathematics
Education*, 33(5).

Kopriva, R.J. (2008). *Improving Testing for English
Language Learners*. Routledge: New York.

Koretz, D., R. Linn, S. Dunbar, and L. Shepard (1991). "The
Effects of High-Stakes Testing on Achievement: Preliminary
Findings About Generalization Across Tests." Paper
presented at the annual meeting of the American Educational
Research Association, Chicago, IL.

Kreig, R.G., & Uyar, B. (2001). Student performance in
business and economics statistics: Does exam structure
matter? *Journal of Economics and Finance*, 25(2), 229-241.

Kulhavey, R. W., Dyer, J. W., & Silver, L. (1975). The
effects of notetaking
and test expectancy on the learning of text material.

- Journal of Educational Research*, 68, 363-365.
- Kulm, G. (1994). *Mathematics Assessment: What Works in the Classroom*. Jossey-Bass: San Francisco.
- Lam, T.C. (1995). Fairness in performance assessment. *Eric Digest*, ED391982, 1-6.
- Lane, S., Silver, E.A., Ankenmann, R.D., Cai, J., Finseth, C., Liu, M., Magone, M.E., Meel, D., Moskal, B., Parke, C.S., Stone, C.A., Wang, N., & Zhu, Y. (1995). *QUASAR Cognitive Assessment Instrument (QCAI)*. Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.
- Lange, B. (1981). Promoting testwiseness. *Journal of reading*, 24(8), 740-743.
- Leiva, M. (1995). Implementing the professional standards for teaching mathematics: empowering teaching through the evaluation process. *Mathematics Teacher*, 88(1), 44-47.
- Linn, M.C., & Hyde, J.S. (1989). Gender, mathematics, and science. *Educational Researcher*, 18(8), 17-19, 22-27.
- Liu, O. L. & Wilson M. (2009). Gender differences in large-scale math assessments: PISA Trend 2000 and 2003. *Applied Measurement in Education*, 22, 164-184.
- Linn, R. (2000). "Assessments and accountability." *Educational Researcher* 29 (2), 4-16.

Lissitz, R. W., & Hou, X. (2007). Multiple-choice items and constructed-response items: Does it matter? Retrieved from the World Wide Web on November 30th, 2010,

<http://www.education.umd.edu>

Livingston, S.A. (2009). Constructed-response questions: Why we use them, how we score them. *ETS: R&D Connections* 11, 1-8.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement*, 31(3), 234-250.

Ma, X., & Willms, D. (1999). Dropping out of advanced mathematics: How much do students and schools contribute to the problem? *Educational Evaluation and Policy Analysis*, 21(4), 65-83.

Martinez, M.E. (1991). A comparison of multiple-choice and constructed figural response items. *Journal of Educational Measurement*, 28, 131-145.

Martiniello, M. (2008). Language and the performance of English Language Learners in math word problems. *Harvard Educational Review*, 78, 333-368.

Mauldin, R.K. (2009). Gendered perceptions of learning and fairness when choice between exam types is offered. *Learning in Higher Education, 10*(3), 253-264.

Marzano, R. J., Pickering,, D., & McTighe, J. (1993). *Assessing Student Outcomes: Performance Assessment using the Dimensions of Learning Model*. Alexandria, VA: ASCD.

Marzano, R., Pickering, D.L., & Pollock, J.E. (2001). *Classroom Instruction That Works: Research-Based Strategies for increasing Student Achievement*. Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. (2003). What works in schools: Translating research into action. ASCD.

Mazzeo, J., Schmitt, A.P., & Bleistein, C.G. (1991). *Do women perform better, relative to men, on constructed-response tests or multiple-choice tests? Evidence from the Advanced Placement Examination*. Paper presented at the annual meeting of National Council of Measurement in Education, Chicago, IL.

McPhail, I. (1981). Why teach testwiseness? *Journal of Reading, 25*(1), 32-38.

Mehrens, W. A. (1998). Consequences of Assessment: What is the Evidence? *Educational Policy Analysis Archives, 6*(13), 1-35.

Menken, K. (2000). What are the critical issues in wide-scale assessment of English language learners? Washington, DC: National Clearinghouse for Bilingual Education. Retrieved November 4, 2002, from <http://www.ncela.gwu.edu/ncbepubs/issuebriefs/ib6.htm>

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, 50, 741-749.

Metcalf, K., & Cruickshank, D. (1991). Can teachers be trained to make clear presentations? *Journal of Educational Research*, 85(2), 107-116.

Millman, J., Bishop, C.H., & Ebel, R. (1965). An analysis of testwiseness. *Educational and Psychological Measurement*, 25, 707-726.

Moskal, B. M. (2000). "Scoring rubrics: What, when and how?" *Practical Assessment, Research & Evaluation*, 7 (3), 1-15.

Myerberg, J. (1996). Performance on different test types by racial/ethnic/ group and gender. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

National Council of Teachers of Mathematics (1987). *How to Evaluate Progress in Problem Solving*. Reston, VA: Author

National Council of Teachers of Mathematics (1991).
Mathematics Assessment: Myths, Models, Good Questions, and Practical Suggestions. Reston, VA: Author.

National Council of Teachers of Mathematics (2000).
Curriculum and Evaluation Standards for School Mathematics. Reston, VA: Author.

National Regional Educational Laboratory (NREL) (1988).
Measuring Thinking in the Classroom. Oak Park, Ill: Author.

National Association of State Boards of Education. (2001).
A primer on state accountability and large-scale assessments. Retrieved from the World Wide Web:
http://www.nasbe.org/Educational_Issues/Reports/Assessment.pdf on 10/01/2010.

Ohlsen, M. (2007). Classroom Assessment Practices of Secondary Mathematics Teachers of NCTM. *American Secondary Education*, 36(1).

Oosterhof, A.C. & Coates, P.K. (1984). Comparison of difficulties and reliability of quantitative word problems in completion and multiple-choice formats. *Applied Psychological Measurement*, 8, 287-294.

Paterson, J. S. (2002). What's in a name? A new hierarchy for question types. Scottish Center for Research into On-Line Learning and Assessment, School of Mathematics and Computer Science.

Pawlak, P. L. (2009). Common Characteristics and Classroom Practices of Effective Teachers of High-Poverty and Diverse Students. (Doctoral dissertation, Claremont Graduate University, 2009).

Paxton, M. (2000). A linguistic perspective on multiple-choice questioning. *Assessment and Evaluation in Higher Education*, 25(2), 109-120.

Pomplun, M. & Capps, L. (1999). Gender Differences for Constructed-response Mathematics Items. *Educational and Psychological Measurement*, 59(4), 597-614.

Popham, W. J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership*, 56(6), 8-15.

Popham, W. J. (2010). Classroom Assessment: What Teachers Need to Know. Pearson Allyn & Bacon.

Powell, J.L, & Gillespie, C. (1990). *Assessment: All Tests Are Not Created Equally*. Paper presented at the Annual Meeting of the American Reading Forum, Sarasota, Florida.

Rafilson, F. (1991). "The case for validity generalization." *Practical Assessment, Research & Evaluation*, 2 (13), 54-67.

Ravitch, D. (2000). *Left Back: A Century of Battles Over School Reform*. Touchstone: New York.

Reeves, D. (2004). *Accountability in Action: A Blueprint for Learning Organizations* (2nd Edition), Denver, CO: Advanced Learning Press.

Rickards, J. P., & Friedman, F. (1978). The encoding versus the external storage hypothesis in note taking. *Contemporary Educational Psychology*, 3, 136-143.

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

Roediger, H. L. III (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology*, 31(5), 1155-1159.

Rogers, W.T., & Harley, D. (1999). An empirical comparison of three-and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement*, 59(2), 234-247.

Rosenshine, B. (1976). Classroom instruction. In N.L.Gage (Ed). *The psychology of teaching methods*. The National Society for the Study of Education Seventy-seventh Yearbook.

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

Scouller, K. M, & Prosser, M. (1994). Students' experiences in studying for multiple choice question examinations. *Studies in Higher Education, 79*(3), 267-279.

Scouller,K. M. (1998). The influence of assessment methods on students' learning approaches: multiple-choice question examinations versus assignment essay. *Higher Education, 35*, 453-472.

Shulman, L. S. (1981). Disciplines of inquiry in education: An overview. *Educational Researcher 10*(6), 5-12, 23.

Simkin, M.G., & Kuechler, W.L. (2005). Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education, 3*(1),73-91.

Snow, R.E. (1993). Construct validity and constructed-response tests, in E.B Randy, and C.W. William (Eds), *Construction versus choice in cognitive measurement*. Hillsdale, New Jersey: Lawrence Erlbaum, 45-60.

Stenmark, J.K (1989). *Assessment Alternatives in Mathematics: An Overview of Assessment Techniques that Promote Learning*. Regents: University of California.

- Stevenson, A., & Stigler, J. (1992). *The Learning Gap*. Touchstone: New York.
- Stone, A. (1998). The Metaphor of Scaffolding: Its Utility for the Field of Learning Disabilities. *Journal of Learning Disabilities*, 3(4), 344-364.
- Stronge, J.H. (2002). *Qualities of Effective Teachers*. Alexandria: Association for Supervision and Curriculum Development.
- Tankersley, K. (2007). *Tests That Teach: Using standardized tests to improve instruction*. Alexandria, VA : Association for Supervision and Curriculum Development.
- Tobias, S. & Duffy, T. (2009) Eds. *Constructivist instruction*. Routledge: New York.
- Traub, R.E & Fisher, CW. (1977). On the equivalence of constructed response and multiple-choice tests. *Applied Psychological Measurement*, 1, 355-369.
- Traub, R.E. & MacRury, K. (1990). Multiple-choice vs. free response in the testing of scholastic achievement. *Tests and Trends*, 8, 128-159.
- Vygotsky, L. S. (1992). *Educational psychology*. St Lucie Press, Florida.
- Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist

theory of test construction. *Applied Measurement in Education* 6, 103-118.

Walstad, W.B., & Becker, W.E. (1994). Achievement differences on multiple-choice and essay tests in economics. *American Economic Review*, 84(2), 193-196.

Wang, J. (1995). *Critical Values of Guessing on True-False and Multiple-Choice Tests*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Wells, C.S. & Wollack, J.A. (2003). An instructor's guide to understanding reliability. Retrieved on June 4th, 2010 from the World Wide Web,

<http://www.testing.wisc.edu/reliability>,

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *The Phi Delta Kappan*, 70(9), 703-713.

Yackel, E. and Cobb, P. (1996). Socio-mathematical norms, argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education*. 27(4), 458-477.

Zeidner, M. (1987). Essay versus multiple choice type classroom exams: The students' perspective. *Journal of Educational Research*, 80(6), 352-358.

Zimmerman, D.W., & Williams, R.H. (2003). A new look at the influence of guessing on the reliability of multiple-choice tests. *Applied Psychological Measurement*, 27(5), 357-371.

Zucker, S. (2003). *Fundamentals of Standardized Testing*.
San Antonio TX: Harcourt Assessment, Inc.