7-18-2005

Relationship between course-taking behavior, gender, and mathematics achievement on the Missouri Assessment Program (MAP)

Geraldine Dressel Baumgart
University of Missouri-St. Louis

RELATIONSHIP BETWEEN COURSE-TAKING BEHAVIOR, GENDER, AND MATHEMATICS ACHIEVEMENT ON THE MISSOURI ASSESSMENT PROGRAM (MAP)

by

GERALDINE DRESSEL BAUMGART

B.S. Mathematics, Maryville University - St. Louis
M.Ed. Secondary Education, University of Missouri - St Louis
M.Ed. Counseling, University of Missouri - St. Louis

A DISSERTATION

Submitted to the Graduate School of the

UNIVERSITY OF MISSOURI - ST. LOUIS
In partial Fulfillment of the Requirements for the Degree

DOCTOR OF EDUCATION

in

EDUCATIONAL ADMINISTRATION

June, 2005

Advisory Committee

Lloyd I. Richardson, Ph.D.
Chairperson
Kathleen Sullivan Brown, Ph.D.
Cody Ding, Ph.D.
Richard Friedlander, Ph.D.
Thomas Schnell, Ph.D.

ACKNOWLEDGEMENTS

I wish to express my gratitude to the people who contributed to the completion of this study. First, I thank my committee members: Dr. Kathleen Sullivan Brown for her constant encouragement and support as my advisor as well as her careful and unbelievably swift editing of numerous drafts, Dr. Cody Ding for patiently sharing his knowledge and understanding of statistics with me, Dr. Richard Friedlander for bringing the perspective of practicing mathematics teachers to the discussion, and Dr. Thomas Schnell for his generous gift of time to painstakingly edit and guide me through multiple revisions of this paper. My special thanks go to my chairperson, Dr. Lloyd Richardson, for helping me define a study that I could be passionate about, for having the patience to stay with this project through the endless derailments that make up my complicated life, and for nudging me to always work a little harder so that the outcome would be something of which I can be proud.

I am also grateful to several other groups of people in the education and research community. The study would not have been possible without the willingness of the participating school district to provide me with access to their data. The university staff in the libraries, computer labs, and various offices cheerfully helped me over numerous hurdles to get to the finish line. Many fellow researchers outside of the university went out of their way to provide me with information about their research that informed my own study. I will show my gratitude to them by being generous to those who ask for my help. My coworkers lifted me up throughout this long process with their support, understanding, and encouragement.

I wish to thank my friends and family for being a source of love, strength, and joy. To my four sons: Joel, Mike, Matt, and Dan, I love you very much and I know you are all

happy this is over. To my grandchildren, Aidan and Nora, you always remind me that people are more important than things. I look forward to spending much more time with both of you. Much love and special thanks to my daughter, Mary, who gave cheerfully of her time to help me with typing as well as going with me to every library in town to find and copy journal articles. Love, gratitude and a big hug to my husband, Steve, who must have been thinking this is the 'worse' that I signed up for many years ago. I could not have done this without your willingness to pick up the slack and cook, drive, help with homework, and then read every version of each chapter.

Finally, I thank God for seeing me to the end of this and I trust in His wisdom to help me use this knowledge to teach and help others.

ABSTRACT

This study examined the relationship between student course taking, specifically the year of Algebra completion (grade 8, 9, 10, or not completed), and performance on the Missouri Assessment Program (MAP) mathematics test in grades 8 and 10. Data collected were student scores on the MAP tests, TerraNova tests in Communication Arts and Mathematics, student math grades, and demographic factors of gender and race. The sample of 512 students was taken from one school district in east central Missouri.

The MAP mathematics tests contain 3 item types. Item type was statistically significant with both males and females scoring highest on Multiple Choice followed by Constructed Response and Performance Event items. Males and females had similar profiles for item types at both grade levels with males performing better than females on each item type at grades 8 and 10. The only statistically significant gender difference was on Multiple Choice items in grade 10.

Course taking was significantly related to performance on the six MAP mathematics content strands. Number Sense, Geometry and Spatial Sense, Data Analysis and Probability, Patterns and Relationships, Mathematical Systems, and Discrete Mathematics organize the MAP content. Number Sense and Mathematical Systems content strands both found males performing significantly better than females. Geometry and Spatial Sense was the only strand that yielded a significant interaction effect of gender by course taking with males gaining significantly in advantage over females as Algebra was completed earlier.

Course taking was significantly related to overall MAP and TerraNova mathematics scores. ANCOVA analyses used TerraNova language scores as a covariate to isolate the

effect of course taking on MAP performance. The ANCOVA employed course taking and gender as independent variables and explained 70% of the variance in MAP 8 scores and 53% of the variance in MAP 10 scores. Both course taking and gender were significant main effects.

A logistic regression analysis revealed significant predictors of MAP 10 mathematics performance to be MAP 8 mathematics performance, Math GPA in grades 8 through 10, gender, and completion of an Algebra course in grade 8. Qualified students should be encouraged to take Algebra in grade 8.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

The current national and state level systems of accountability for public education require educators to be more vigilant than ever in monitoring the alignment between the curriculum that students experience and the assessments that are used to measure student achievement. The relative merits of various measures of student achievement spark many of the debates in education today. The literature points to measures that include standardized test scores, grade point average, and performance on the criteria being assessed (Alexander & Pallas, 1984; Roth, Crans, & Carter, 2001). Researchers consistently report that student course-taking behavior in mathematics is strongly and positively correlated with performance on mathematics assessments (Alexander & Pallas, 1984; Bohr, 1994; Jones, Davenport, Bryson, Bekhuis, & Zwick, 1986; Sebring, 1985; Smith, 1996; Useem, 1990). Gender studies have shown that the gender gap in mathematics, favoring males, is narrowing; however, the gender gap, favoring females, in reading and writing persists (Coley, 2001; Fan & Chen, 1997; Gambell & Hunter, 1999; Han & Hoover, 1994; Hedges & Nowell, 1995; Hyde, Fennema, & Lamon, 1990; Kleinfeld, 1998; Lee & Ware, 1986; Maccoby & Jacklin, 1974; McLure, 1998; Nowell & Hedges, 1998; Pallas & Alexander, 1983; Pomplun & Sundbye, 1999; Rebhorn & Miles, 1999; Ryan & Fan, 1996; Taylor, Leder, Pollard, & Atkins, 1996; Wainer & Steinberg, 1992; Wilder & Powell, 1989; Willingham & Cole, 1997).

This dissertation examined the relationship between two different measures of student achievement in mathematics and the course-taking behaviors of students, taking into account gender. Included in the student achievement data were individual student scores on the mathematics portion of the Missouri Assessment Program (MAP) tests in

grades 8 and 10, and the TerraNova Multiple Assessments in grades 8 and 9. The construct of student achievement in mathematics is what both the MAP and the TerraNova mathematics tests purport to measure. The MAP is a criterion-referenced test with multiple choice (MC), constructed response (CR), and performance event (PE) items. The Missouri Department of Elementary and Secondary Education (DESE) contracts with CTB McGraw-Hill to provide the MC section of the MAP, which is the Survey portion of the TerraNova. The TerraNova Multiple Assessment is a norm-referenced test with both MC and CR items. Because the MAP test is intended to measure what students know and are able to do, the students are required to provide written responses to open-ended questions. Researchers have found a positive relationship between achievement levels in reading and mathematics (Abedi, Lord, & Hofstetter, 2001; Czujko & Bernstein, 1989). To examine student ability in reading and writing, student scores on the Communication Arts portion of the TerraNova were included in the data analysis.

Background

For decades, legislators, educational policymakers, and the general public have voiced concern over the relatively poor performance of American students. In 1957, the Soviet launch of the satellite, Sputnik, led to what has been referred to as "post-Sputnik hysteria on the parts of educational leaders" (Renzulli, 2004, p. 5). Americans felt that this single event represented a threat to our national security because the Soviet Union had pulled ahead of the United States in the “space-race.” In response to Sputnik, the United States Congress rushed to pass the National Defense Education Act (NDEA) in

1958. The NDEA provided federal funds for many educational programs including student loans, high school guidance counselors, graduate fellowships, and aid for improving teaching in mathematics, science, and languages (http://ishi.lib.berkeley.edu/cshe/ndea/ndea.html).

In 1965, as part of President Lyndon Johnson's "Great Society," the 89th Congress of the United States passed the Elementary and Secondary Education Act (ESEA). The main intent of ESEA was to provide funding for programs that would allow equal access to a quality education for all elementary and secondary students in America, especially those students who were most disadvantaged. Most of the $1 billion per year allocation from ESEA went to Title I, which funded programs to meet the needs of children from low-income families. This financial assistance was distributed to 90% of all schools in the country, including non-public schools. Senator Robert Kennedy was among legislators who wanted a means for evaluating the Title I programs in order to hold schools accountable (Carleton, 2002). The Title I Evaluation and Reporting System (TIERS) was developed for this purpose. The use of the Normal Curve Equivalent (NCE) scale for reporting test scores became prevalent with the TIERS program (Linn, 2000). While government funding for education was on the rise, increased public attention was being focused on ways to measure the effectiveness of our educational system.

Work began in the 1960s on planning and developing a national student assessment system. In 1969, the National Assessment of Educational Progress (NAEP) began. At the outset, policymakers in Washington, DC promised that there would be no data reported at the state, district, or student level. This promise was necessary to gain support from educators and state and local officials who feared too much federal

influence would ruin our educational system of local control (Vinovskis, 1998). This early assessment eventually evolved into the two separate NAEP systems we have today. The second NAEP system would not be fully developed until the early 1990s.

In the 1970s, many states, including Missouri, satisfied the need for accountability at the state and local level with Minimum Competency Testing (MCT). As the name implies, these tests were designed to assess a student's mastery of a prescribed list of basic skills. In some states, these tests were part of a system of graduation requirements. Since the competencies were at such a low level, these MCTs did nothing to raise the level of expectations for the academic content that students were learning. There was still widespread concern that students in the United States would not be ready for the 21st century workplace demands.

This concern led President Reagan's Secretary of Education, Terrell Bell, to create the National Commission on Excellence in Education (NCEE) on August 26, 1981. The task he gave to the commission was to assess the condition of American education and report to him within 18 months with findings and recommendations. *A Nation at Risk: The Imperative for Educational Reform* (National Commission on Excellence in Education, 1983) became one of the most significant educational reports in the last half of the 20th century. The NCEE felt that American curriculum lacked the rigor necessary to prepare our students to compete in a global economy. The commission's finding related to assessment in the US indicated that "'Minimum competency' examinations (now required in 37 States) fall short of what is needed, as the 'minimum' tends to become the 'maximum,' thus lowering educational standards for all" (National Commission on Excellence in Education, 1983, Findings Section, Findings regarding

expectations Subsection). This report led to changes in curriculum, instruction, and graduation requirements across the United States.

The call for higher expectations for all students, and the need to prepare students for the mathematical, scientific and technological demands of the future continued. Mathematics educators responded by developing content standards by the end of the 1980s. The National Council for Teachers of Mathematics (NCTM) published the *Curriculum and Evaluation Standards for School Mathematics* (National Council of Teachers of Mathematics, 1989). These were closely followed by *Professional Standards for Teaching Mathematics* (National Council of Teachers of Mathematics, 1991), as well as *Assessment Standards for School Mathematics* (National Council of Teachers of Mathematics, 1995). These NCTM Standards would provide a basis for most of the standards and frameworks in mathematics developed by the individual states, including the Missouri Show-Me Standards.

In 1989, "President George H.W. Bush convened the first National Education Summit to discuss national educational goals with state governors" (National Research Council, 2002, 19). The National Education Goals Panel (NEGP) was formed as a result of this first summit meeting. Later, President Clinton continued the work that was begun by President Bush and in 1994 the United States Congress passed two important pieces of educational reform legislation. Goals 2000: Educate America Act of 1994 became law on March 31, 1994. It stated, among other things, “by the year 2000 United States students would be first in the world in mathematics and science achievement" ("Goals 2000," 1994). Throughout the history of education-related federal legislation, lawmakers have carefully avoided the imposition of a national curriculum or the requirement of a national

assessment. Federal lawmakers consistently state that control of our public schools should remain a state government responsibility. Goals 2000 stopped short of requiring a national curriculum or a national assessment; rather, it outlined national goals and gave states flexibility in designing programs to achieve those goals (Carleton, 2002).

On October 20, 1994, the Improving America's Schools Act (IASA), amended ESEA and the accountability requirements related to the receipt of Title I funds. There was now a requirement that annual assessments be given in reading or language arts and mathematics and that these assessments be aligned to "challenging content and performance standards" and be administered at least once during grades 3-5, 6-9, and 10-12 ("IASA," 1994). These new requirements included (a) the identification of levels of student performance, such as "proficient" or "advanced;" (b) the reporting of disaggregated scores (by gender, by each major racial and ethnic group, by English proficiency status, by students with disabilities, and by economically disadvantaged students); (c) the reporting of scores at the state, district, and building levels; and (d) the charting of "adequate yearly progress." At the time IASA was passed, many states did not have a system in place for these assessment and accountability requirements, so they were given until 2001 to comply or adopt an existing accountability system from another state (Reckase, 1999).

On the national assessment front, amidst great disagreements between educators and lawmakers, NAEP was expanded in the 1990s from a strict MC test to one that included constructed response items and allowed students to use calculators and other materials while taking the test. With NAEP's expansion, state-level data would now be collected. The NAEP does not compare our students to those from other nations; however,

politicians and the public consistently monitor our students progress on NAEP as well as U.S. students' performance relative to students from other parts of the world. The poor performance of United States mathematics students on international assessments was cited as part of the rationale for forming the commission that wrote *A Nation at Risk.* Despite educator's efforts to implement the recommendations of the commission, in 1995, the Third International Mathematics and Science Study (TIMSS) also found the US fared poorly by comparison to the other 21 industrialized nations (Haury & Milbourne, 1999).

Many researchers have proposed various analyses to account for the disappointing performance of American mathematics students on TIMSS and other international mathematics assessments that preceded it. Frequently, the mathematics curriculum is considered at least partially to blame (McKnight et al., 1987; Metcalf, 2002; U.S. Department of Education, National Center for Education Statistics, 1996). Of particular concern is the practice in U.S. schools of tracking students early in their academic careers, before they reach high school. From tracking concerns come concerns about equity and access for all students. There is a significant body of research that documents that the opportunity to learn important mathematics is not the same for all American students (Gamoran, 1987; McKnight et al., 1987; Moses & Cobb, 2001; Oakes, Ormseth, Bell, & Camp, 1990; Useem, 1990; Ware, Richardson, & Kim, 2000).

In *A Nation at Risk*, the NCEE recommendation for graduation requirements was that all students would take the *Five New Basics.* These new basics were said to form the core of the modern curriculum. This core included 4 years of English; 3 years each of mathematics, science, and social studies; and 1/2 year of computer science. The recommendation for college bound students included 2 years of foreign language. In 1982

only 2% of public high school graduates took the college prep core. By 2000, the percent of students nationally taking the college prep core had increased to 31%, the percent of females completing the curriculum was 33.2% vs. 28.6% males. The average number of Carnegie units earned by public high school graduates in mathematics (Algebra or higher) went from 1.74 in 1982 to 2.95 in 2000. The scales tipped in favor of girls taking more Carnegie units in mathematics from 1982 to 2000. Boys took 1.77 to girls' 1.71 credits in 1982, but in 2000 girls took 3.03 to boys' 2.86 credits (National Center for Education Statistics, 2003).

The percent of students in grades 9-12 who are taking mathematics at a level of Geometry or higher is at 48% nationally, which is an increase of 14% since 1990. In Missouri, 55% of high school students are taking courses at these levels, an increase of 19% from 1990. The nation reports 89% of high school students are enrolled in a mathematics course (any mathematics course), while 95% of Missouri high school students are enrolled in a mathematics course (Blank & Langesen, 2003).

In more than 40 states, high school graduation requirements have increased over the last two decades in response to pressures from the public (Blank & Langesen, 2003). At the time *A Nation at Risk* was published 35 states required only 1 year of mathematics for graduation; by 2002, 27 states required 3 credits of mathematics (National Center for Education Statistics, 2002a).

Of the 37 jurisdictions reporting graduation requirements and mathematics course enrollments in 2002, Missouri is 1 of only 7 states that still required only 2 credits in mathematics for graduation. However, Missouri has the highest percentage (89%) of students taking Algebra II or Integrated Mathematics 3 by graduation. This is an increase of

31% from 1990 (Blank & Langesen, 2003). While Missouri students are increasing the intensity and number of mathematics courses taken in grades seven through 12, the average Missouri scores on the NAEP are not very different from the national average scores.

While attention has been focused on our performance as a nation and as states, the federal laws continue to place the responsibility for defining standards and monitoring mastery at the level of state departments of education. Recognizing the need for legislation mandating standards and assessments at the state level, in 1993 Missouri passed The Outstanding Schools Act, which included a provision for an assessment system that would measure what Missouri students "know and are able to do" relative to performance standards. The Show-Me Standards for process and content were developed in Missouri as part of the mandate of the Outstanding Schools Act. There are 73 standards; 33 performance (process) standards and 40 knowledge (content) standards, six of these content standards are in mathematics. The mathematics content standards are divided into six categories of Number Sense, Geometry and Spatial Sense, Data Analysis and Probability, Patterns and Relationships, Mathematical Systems, and Discrete Mathematics. These complete Show-Me Standards, including process and content, are also available on the web at http://dese.mo.gov/standards/mathematics.html.

The mathematics content standards defined the strands that make up the organizational structure of the Framework for Curriculum Development in Mathematics, K-12 (Missouri Department of Elementary and Secondary Education, 1996). The frameworks were not intended to be a state curriculum. They were intended to give Missouri's educators information about what students should know and be able to do in

grades K-4, 5-8, and 9-12. This document included sample learning activities to aid educators in providing instruction consistent with the intent of the frameworks. The *Supplement to the Curriculum Frameworks, Mathematics K-12* (Missouri Department of Elementary and Secondary Education, 2001) was developed to give educators more activities organized within each of the content strands. Again the purpose of the document was to help school district curriculum specialists and classroom teachers structure lessons and assessments that would lead to student mastery of the Show-Me Standards (http://dese.mo.gov/divimprove//curriculum/frameworks supplement/ math1.html).

The Missouri Assessment Program (MAP) was developed as a criterion-referenced, performance-based assessment system used to measure student progress toward mastery of the Show-Me Standards. On April 27, 1995, the Missouri State Board of Education adopted a policy statement regarding the four main purposes of the MAP. They designated the purposes of the assessment program as "(1) improving students' acquisition of important knowledge, skills, and competencies; (2) monitoring the performance of Missouri's educational system; (3) empowering students and their families to improve their educational prospects; and (4) supporting the teaching and learning process" (Bartman, 1998).

In addition, the board explained that it would serve these purposes by providing data that could be used to make informed educational judgments "concerning individual students, groups of students, and educational programs" (Bartman, 1998, p. 2). They also identified "three major uses of assessment results: instructional, guidance and counseling, and administrative" (p. 1). The present research study was carried out in the context of

these intended uses of the MAP. In order to make judgments about curriculum and assessment alignment, it is important to analyze the relationship between student course-taking behavior and MAP scores.

The MAP mathematics test was the first content test completed. There were field and pilot tests but the first year that all Missouri public school students were required to take the MAP mathematics tests in grades 4, 8, and 10 was in 1998. The MAP contains three sessions and three types of items. The entire test (all three sessions) takes approximately 3 to 5 hours to complete. Only the Multiple Choice session is timed. The Missouri Department of Education contracted with CTB McGraw-Hill to support the development and administration of the MAP. The Multiple Choice (MC) component is the Survey portion of the TerraNova, a nationally norm-referenced achievement test published by CTB McGraw-Hill. The Constructed Response (CR) items require students to supply an answer and in some cases to show their work and explain. The Performance Event (PE) items not only measure students' knowledge but also their ability to apply that knowledge to complex real-life situations. Students are expected to work through a multi-step process, justify their solution, and provide explanations that include showing and labeling their work. These (performance events) are more complex problems and there can be multiple acceptable approaches to a correct answer. The MC portion is machine scored and the CR and PE portions are hand-scored by hired raters who have been trained to read and score such items, using scoring guides (or rubrics).

> Rigorous training and the use of item-specific scoring guides ensure uniformity of scoring. Scoring is organized and conducted by Missouri's contractor CTB/McGraw-Hill. The Department will monitor the reliabil-

> ity and validity of each subject area scoring by organizing groups of Missouri teachers to re-score a representative sample of student responses. (Bartman, 1998, p. 7)

Information on reliability and validity was supplied by a staff member at the Missouri Department of Elementary and Secondary Education (W. Gerling, personal communication, July 6, 2003) and the same information is available on the web at the following address: www.dese.state.mo.us/divimprove/fedprog/discretionarygrants/ReadingFirst/DMAP.pdf The interrater reliability for the 1999 and 2000 MAP assessments showed the median percent of perfect agreement for two readers of open-ended (CR or PE) MAP mathematics items ranged from 84.39 to 96.04. The percent of adjacent agreement (where two scorers differed by one point) on the 1999 grade 10 MAP mathematics test ranged from 92% to 100%, with a median percent equal to 98%. It is not stated whether those grade 10 figures included CR items. The CR items have possible scores from 0 to 2 and PE items have possible scores from 0 to 4. It is likely that trained CR raters would be within one point almost all of the time. The report also contained statements about validity studies that the Missouri Department of Elementary and Secondary Education (DESE) and CTB McGraw-Hill continue to conduct, but no data were provided. Reliability coefficients were all in the 0.90s for the MAP 8 and MAP 10 mathematics tests.

A critically important part of any standards-based assessment program is the process for defining achievement levels. DESE and CTB McGraw-Hill used the "bookmark procedure" to set the five achievement levels.

> A panel composed of 40 to 45 teachers, parents, and business professionals reviewed the rank ordered test items from field-testing of the MAP. Test items were rank ordered from easiest to the most difficult based upon student performance during the field-test. The panelists placed a bookmark at the point that they thought a student performing at Advanced, Proficient, Nearing Proficient, or Progressing would perform. The panelists then discussed the rationale for their judgments. The judgments of the panel members were averaged to establish cut off points for each achievement level. (Bratberg, 2002, p.11)

The achievement level range of points for each level at grades 8 and 10 follows:

Grade 8 Mathematics:

Advanced: MAP score range 785-915

Proficient: MAP score range 744-784

Nearing Proficient: MAP score range 708-743

Progressing: MAP score range 668-707

Step 1: MAP score range 541-667

Grade 10 Mathematics

Advanced: MAP score range 832-979

Proficient: MAP score range 784-831

Nearing Proficiency: MAP score range 743-783

Progressing: MAP score range 701-742

Step 1: MAP score range 581-700

www.dese.mo.gov/divimprove/assess/GIR_2003.pdf

The website for the unabbreviated MAP mathematics achievement levels is: http://www.dese.state.mo.us/divimprove/assess/Descriptors/Unabbreviated/Mathematics.doc.

The Outstanding Schools Act also requires school districts to develop a comprehensive, board-approved assessment program for all students preschool through 12th grade. "Districts are expected to use MAP results, in conjunction with other indicators, to appraise and strengthen their educational programs" (Bartman, 1998, p.8). Testing in "non-MAP" years is left to the discretion of the local school districts. The district that participated in this study formed a K-12 assessment committee. The committee recommended to the local board of education that the district administer the TerraNova Multiple Assessments in these non-MAP years as one of the other indicators of student performance. The decision was based in large part on the fact that the TerraNova is developed by CTB McGraw-Hill, the test publisher that contracts with Missouri to aid in development, administration, and scoring of the MAP. The TerraNova Multiple Assessments are norm-referenced standardized tests that include both MC and CR items. A staff member at CTB McGraw-Hill supplied information on the relationship between the various forms of the TerraNova test:

> All *TerraNova* configurations (Multiple Assessments, Complete Battery, and Survey) are interconnected and are tied to a common scale. Because of this interconnection, test configurations can vary from grade to grade and year to year and still provide consistent and comparable information. The change in Scale Scores, no matter which version is administered, can

be used to look at growth in achievement from a norm-referenced perspective.

(Sheryl Cole, Evaluation Consultant, email correspondence, June 28, 2004)

Student performance on both the TerraNova and the MAP was analyzed in this study. Researchers have found moderate to strong relationships between student performance on norm-referenced tests and criterion-referenced performance tests (Behuniak & Tucker, 1992; Oescher, Kirby, & Paradise, 1992; Visintainer, 2002). Researchers have recommended the study of the relationship between student performance on these different types of test items because a great deal of academic time is devoted to the administration of performance tests like the MAP (Behuniak & Tucker, 1992; Lukhele, Thissen, & Wainer, 1994; Oescher et al., 1992; Pearson & Garavaglia, 2003; Visintainer, 2002).

Another mandate of Missouri's Outstanding Schools Act is the Missouri School Improvement Program (MSIP). MSIP provides the structure for reviewing and accrediting the 524 school districts in Missouri within a five-year review cycle. The review process includes the areas of resource, process and performance. DESE is now in the third cycle of MSIP reviews. The Standards and Indicators for district accreditation are modified and updated with each new cycle. DESE stated that the "primary focus of the third cycle Process Standards is on improving student performance" (http://dese.mo.gov/divimprove/sia/msip/ThirdCycleFAQ.pdf).

Student course-taking behavior, ACT scores, and student MAP scores are critical to school district accreditation because they contribute to a district's scores on the performance matrix of the MSIP review. DESE identifies specific courses that qualify in

the area of Advanced Courses. The following mathematics classes from DESE's list of advanced courses are offered by the district participating in this study: Geometry, Algebra-Trigonometry, College Algebra, Pre-Calculus, Statistics, and Calculus. In addition to contributing to a district's MSIP review, ACT has determined that successful completion of these and other advanced courses contributes to higher ACT scores and greater performance in college (www.act.org/news/releases/2003/8-20-03.html). Perhaps the most important factor related to students' course-taking behavior is the contribution that rigorous course taking makes to individual student's academic and personal goals.

Just as policymakers, educators, and the public in general have disagreed about how to assess student learning, how to evaluate educational programs, and how to report the data, they also continually debate curriculum issues. Philosophical questions about knowledge have been asked throughout history. What counts as knowledge? Who will have access to this knowledge? These questions are at the heart of conversations about setting standards, designing assessments, tracking, ability grouping, and opportunity to learn (OTL).

The Show-Me Standards have determined what counts as knowledge in Missouri, but is that the same knowledge colleges require? Is it the same knowledge required to be successful in a career immediately after high school? Berlak et al. (1992) contend that test developers begin with an assumption that they have identified a construct (in the case of the MAP mathematics test, the construct would be what counts as mathematics knowledge). Then, they contend, test developers go on to assume that the construct can be measured, that it is somehow quantifiable.

In a presentation in 1997, Kilpatrick made the point that there is no reason to expect that standards-based assessments, regardless of their design can successfully accomplish their goals.

> When curricula have different goals, they can be compared either on the goals they share in common, in which case important things are not measured, or on the entire set of goals, in which case each curriculum is at a disadvantage on the goals it did not attempt… Legitimate comparisons can only be made on common goals, which necessarily fail to capture much of what makes each curriculum unique… If we want to know what mathematics our students are learning from the programs they are in, we need to use instruments that are sensitive to all facets of those programs. (Kilpatrick, 1997, pp. 5-6)

In the case of Missouri's MAP, it is not clear from available reports what the relationship is between curriculum and assessment. It is clear that Missouri has percentages of students close to the national average taking Algebra in grade 8 and taking the ACT core in high school. The percent of Missouri public school students scoring at the Advanced and Proficient levels on the MAP is lower than the percent of students in the accelerated courses. It may be that most of the students with the scores in the top two levels are those in the accelerated track. However, if that is the case then this point from Clune (1998) is well taken: “the fact that the same historically small group of students is still succeeding in the academic fast track does not diminish the need for major advances by other students" (p. 149). Researchers have recommended further study of the relationships between individual student course taking and achievement (Horn, 1990).

The MAP is the instrument for assessing performance of students, schools and districts in Missouri. Curriculum is something that can be modified. It is important to know what the alignment is between student course-taking behavior and achievement on the MAP.

The most recent legislative accountability development at the federal level is once again aimed at mandating the use of state assessment to measure mastery of state standards. On January 8, 2002, President George W. Bush signed the No Child Left Behind Act of 2001 ("NCLB," 2002), a reauthorization of ESEA. Like previous legislation, NCLB also requires states to develop frameworks and assessments and to show adequate yearly progress for all students and all subpopulations of students. NCLB also requires a "report card" from each state documenting progress toward mastery of each state's standards. The difference is that, unlike ESEA, IASA, or Goals 2000, NCLB mandates that all students will be "proficient" in reading and mathematics by 2014. Effective June 10, 2003 all fifty states, the District of Columbia and Puerto Rico had NCLB plans that were approved by the U.S. Department of Education (http://www.ed.gov/new/pressreleases/2003/06/06/02003.html)

NCLB allows states latitude in determining what their state standards and assessments will be, as well as determining strategies and incremental timelines in reaching 100% proficiency by 2014. In Missouri, the decision was made to continue to use the current MAP tests in mathematics at grades 4, 8, and 10. New mathematics tests will be developed for grades 3, 5, 6, and 7 since NCLB mandates testing in each of grades 3 through 8 in mathematics (and communication arts) as well as once during high school for each of these disciplines. NCLB requires states to monitor Adequate Yearly Progress (AYP) toward achieving proficiency for all students as well as all subgroups

defined by each major racial and ethnic group (White, Asian, Black, Hispanic, American Indian, Pacific Islander); by English proficiency status (LEP); by students with disabilities (IEP, IAP); and by economically disadvantaged students (Free or Reduced Lunch status). Missouri set the annual goals for the percent of each group that is proficient or better in mathematics as set forth in Table 1.

Table 1

Missouri's annual goals for the percent of students scoring Proficient or above in Mathematics for the years 2002-2014

2002	'03	'04	'05	'06	'07	'08	'09	'10	'11	'12	'13	'14
8.3	9.3	10.3	17.5	32.1	33.1	54.2	55.2	56.2	77.1	78.1	79.1	100

The Missouri goals for 2005 were recently modified with permission of the U.S. Department of Education. The goal for mathematics for 2005 has been changed from 31.1% to 17.5%. http://dese.mo.gov/news/2005/ayp.htm. Since the percent of Missouri students in grade 8 scoring proficient or above has never exceeded 16.1% and in grade 10 it has never exceeded 12.7%, it is imperative that educators determine what practices will foster higher levels of achievement on the MAP mathematics test.

Both Missouri's Outstanding Schools Act (1993) and the federal government's No Child Left Behind (NCLB, 2002) require states to report scores at the state, district, and building levels. These reports are to include the identification of levels of student performance, such as "proficient" or "advanced;" as well as the reporting of disaggregated scores (by gender, by each major racial and ethnic group, by English proficiency status, by students with disabilities, and by economically disadvantaged students).

The percent of students scoring Proficient or Advanced on the MAP mathematics tests in grades 8 and 10 (Table 2) is still not close to the percent of students who participate in an accelerated curriculum in mathematics, beginning with Algebra in grade 8.

Table 2

Percent of students scoring Proficient or Advanced in MAP mathematics

Participating District Grade 8						Missouri Grade 8					
1998	1999	2000	2001	2002	2003	1998	1999	2000	2001	2002	2003
18.8	17.8	15.5	17.2	22.1	15.7	16.1	10.4	14.1	14.7	13.7	13.9
Participating District Grade 10						Missouri Grade 10					
1998	1999	2000	2001	2002	2003	1998	1999	2000	2001	2002	2003
4.9	14.0	12.6	15.4	16.1	16.0	8.5	9.7	10.3	12.7	10.7	12.4

Despite having an increasing percentage of students in Algebra in grade 8 statewide, as well as a higher percentage of students completing the NCEE recommended core or more, Missouri is not steadily increasing the percentage of students who score at the desirable levels on the MAP. One of the mandates of NCLB is that states must participate biennially in state NAEP beginning in 2002-2003. State level NAEP data "will enable policymakers to examine the relative rigor of state standards and assessments against a common metric" (http://www.ed.gov/admins/lead/account/nclbreference/reference.pdf, p.17-18). Twenty-one percent of Missouri eighth graders scored "Proficient" on the NAEP in 2001 and 28% in 2003. In response to that information, the Missouri Department of Elementary and Secondary Education (DESE) issued a news release, dated November 13 2003, in which they quoted Missouri's Commissioner of Education, Dr. King:

> The NAEP assessment is challenging for kids. It has a structure and expectations that are similar to Missouri's MAP tests. The proficiency scores on both exams are similar, so we believe the NAEP scores offer an important verification that we are 'on track' with our state testing standards. (http://dese.mo.gov/news/2003/naepscores.htm)

Statement of the Problem

Although many factors are known to influence student achievement, some factors can be manipulated and others cannot. Such factors as gender and socioeconomic status are fixed; however, curriculum and instruction are factors that can be manipulated and they are factors that educators constantly strive to optimize. The purpose of this study was to examine the relationship between course-taking behavior in mathematics and performance on the standards-based test in Missouri. While students have been encouraged to take rigorous mathematics courses and take mathematics each year, have these courses prepared students for this standards-based assessment? Although the Missouri Department of Elementary and Secondary Education (DESE) reports disaggregated data at the state, district and building level, there are no data reported that link scores to course-taking behaviors. Is an Algebra course in eighth grade necessary for a student to score at the Proficient or Advanced level on the MAP? If it is necessary, is it sufficient? How does reading ability relate to student performance on the MAP mathematics test? The accountability of No Child Left Behind (NCLB) mandates that schools in the United States must do whatever it takes to help all children become proficient in Mathematics by 2014 ("NCLB," 2002).

Generally, students with greater interest and ability in mathematics choose more rigorous mathematics courses and persist longer in the mathematics pipeline, thus taking more mathematics courses by the end of grade 12. In the school district participating in this study, about 20-25% of eighth grade students took Algebra in eighth grade. A similar pattern exists across the state and nation. However, the total percent of students scoring in the Proficient or Advanced levels on the MAP is considerably lower than the percentage taking the accelerated curriculum. It is not known what the relationship is between student scores and coursework. It is also not known what interaction effects may exist between gender and course taking.

The Missouri Department of Elementary and Secondary Education strives to develop an assessment that is free from bias. However, specific items on the MAP are not available for analysis unless they are released items that will no longer be used. Researchers have documented differential performance by gender on various item types (Anderson, 2002; Bielinski & Davison, 1998; Burton, 1996; Lane, Wang, & Magone, 1996; Muthen et al., 1995; Myerberg, 1996). Researchers have also found that student performance on questions within a content strand may be related to gender and/or course taking (Beller & Gafni, 1996; Bevan, 2001; Bielinski & Davison, 1998; Brosnan, 1998; Harris & Kerby, 1997; Metcalf, 2002). The MAP data at the state level are disaggregated by gender. These data indicate that across the state males outperform females in grades 4, 8, and 10 on the MAP mathematics test with the biggest gap between the genders in grade 10. The 2003 grade 10 MAP mathematics data show that 11.4% of the females scored in the top two levels, while 13.4% of the males scored Proficient or Advanced. National and state level data indicate that girls have surpassed boys in the number of

mathematics courses taken as well as in the completion of the recommended core curriculum. It is not known what the interaction effect is between course-taking behavior and test scores for girls' achievement on the MAP. Do girls who take an accelerated curriculum score the same as boys who take the accelerated curriculum? Is the accelerated curriculum a necessary factor for girls but not for boys?

A factor that may affect performance on the MAP mathematics test is verbal ability. Does verbal ability account for much of the variance in boys' or girls' scores? State level verbal scores show that girls have consistently outperformed boys on the MAP in Communication Arts (CART) in Missouri. The disaggregated data for 2003 in MAP Communication Arts (CART) indicate that females outscore males at a far more dramatic rate than the rates favoring males in mathematics. Females had 38.5%, 38.9% and 27.4% score in the top two levels on MAP CART in grades 3, 7, and 11 respectively. Males had 29.9%, 26.2%, and 16.3% score in the top two levels in grades 3, 7, and 11 respectively. Females' verbal ability may play a significant role in their MAP mathematics performance since the performance on the MAP requires more reading and writing skill than traditional multiple choice tests. When students take the MAP mathematics test they are being asked to justify their solutions and explain the procedures used to solve problems; verbal skills and language proficiency may be highly correlated with MAP mathematics performance (Czujko & Bernstein, 1989; Pearson & Garavaglia, 2003; Visintainer, 2002).

Researchers frequently use Item Response Theory methods to determine whether there is differential performance on specific test items (Hamilton, 1999; Rock & Pollack, 1995b). In the case of the MAP test, the points received by the student for a particular

item are reported. However, it is not possible to see the actual items nor is it possible to analyze the student's actual responses because the questions and individual student responses are not returned to the school districts after the test. In addition, the method used by DESE to arrive at a scale score for the MAP is not made public. Questions are given various weights but the information about the weighting is not released. The original benchmark scores that were set in 1998 have remained the same (see page 13).

Although students in an accelerated mathematics curriculum are likely to have higher interest and greater ability in mathematics, it is unknown how a course in Algebra I in grade 8 influences performance on the MAP test. The grade 8 Algebra I curriculum in the district participating in this study is a standard Algebra I curriculum, identical to the one used in the high school. Much instructional time is devoted to factoring, coordinate graphing, linear equations, and simplifying expressions. Many of these topics may fall within the Show-Me Standard content strand of Mathematical Systems. The MAP assesses topics that fall in six strands. It is unclear whether time spent on algebra topics takes away time that might be devoted to mastery of these other MAP topics. The district participating in this study has completed an analysis of the estimated proportion of instructional time devoted to each strand in the Algebra I curriculum in grade 8, with the understanding that topics frequently overlap in mathematics. The percentage of instructional time spent on each strand are estimated to be 16.4% for Number Sense, 10.1% for Geometry and Spatial Sense, 6.8% for Data Analysis, 10.2% for Patterns and Relationships, 48.2% for Mathematical Systems, and 7.9% for Discrete Mathematics. To spend less time on Mathematical Systems might improve MAP scores on the other strands but at the same time it might do a disservice to students who are

preparing to take higher level courses in mathematics and need the instructional time devoted to learning algebra.

Purpose of the Study

The primary purpose of the study is to examine the relationship between student achievement scores on the Missouri Assessment Program (MAP) mathematics tests in grades 8 and 10 and mathematics course-taking behavior, especially the year of Algebra completion. The interaction effect of course taking and gender was also examined. The data used in this study were examined using regression and correlation techniques to describe relationships and to determine if a formula could be developed to predict performance on the MAP in grade 10. The subjects were students who were tested with each measure and enrolled in grade 10 in 1999-2000, 2000-2001, 2001-2002, or 2002-2003 in a suburban school system in Missouri.

Research questions

The following research questions will be addressed by this study.

1. Is there a relationship between gender and item type on the grade 8 or grade 10 MAP mathematics tests? The MAP contains three item types: Multiple Choice (MC), Constructed Response (CR), and Performance Event (PE).

2. Is there a relationship between course taking, gender and content strand scores on the eighth grade MAP? (The MAP contains questions on six content strands: Number Sense, Geometry/Spatial Sense, Data Analysis/Probability, Mathematical Systems, and Discrete Mathematics.)

3. Is there a relationship between scores on the eighth-grade or the tenth-grade Missouri Assessment Program (MAP) test in mathematics and course-taking behavior in mathematics, taking into account gender?
4. Can the proficiency level(s) on the tenth-grade Missouri Assessment Program (MAP) be predicted by some combination of the factors of mathematics course taking, performance on eighth-grade MAP mathematics test, TerraNova Multiple Assessments in Communication Arts, student grade point average (GPA) in mathematics for grades 8 through 10, race, or gender?
5. Is there a relationship between scores on the TerraNova mathematics test in grades 8, 9, and 10 and course-taking behavior in mathematics?

Definition of Terms

The following terms are used throughout the study and require definition.

Accountability refers to the use of test scores to judge the relative success or failure of schools (National Research Council, 2002)

Constructed response is a short answer question on the MAP. Students are not limited to a choice of options, as they are with Multiple Choice questions.

Criterion-referenced test is one in which scores are determined by comparing performance to a pre-specified standard (criterion) (Gall, Borg, & Gall, 1996).

High-stakes tests are "tests used as direct measures of accountability for students, educators, schools, or school districts, with significant sanctions or reward attached to test results" (Gordon & Reese, 1997, p. 345). In Missouri the stakes for school districts are points on the performance matrix of the Missouri School Improvement Plan (MSIP). MSIP is the program that evaluates public school districts for accreditation in Missouri.

Missouri Assessment Program (MAP) is a state-mandated performance-based assessment system for use by all public schools in the state, as required by the Outstanding Schools Act of 1993. The assessment system is designed to measure student progress toward

meeting the Show-Me Standards, 73 rigorous academic standards that were adopted by the State Board of Education in January 1996.

Three types of items are used on the tests to evaluate student achievement: the familiar multiple-choice questions that require students to select the correct answer; short-answer, constructed-response items that require students to supply (rather than select) an appropriate response; and performance events that require students to work through more complicated problems or issues.
(http://services.dese.state.mo.us/divimprove/assess/general.html)
The mathematics portion of the MAP is administered in grades 4, 8 and 10.

Missouri School Improvement Plan (MSIP) is the program that evaluates public school districts for accreditation in Missouri.

Norm-referenced tests are standardized tests that have been given to a sample group for the purpose of creating a comparison group. Student scores are compared to the scores of the normative sample usually as percentile rank scores.

Show-me standards are those set out by the state department of education in Missouri. There are Performance (process) standards and Knowledge (content) standards. The knowledge standards (in mathematics) are based on the NCTM standards. The Show-Me "standards serve as a blueprint from which local school districts may write challenging curriculum to help all students achieve their maximum potential" (Missouri Department of Elementary and Secondary Education, 1996).

Delimitations

The sample for this study will include only those students who were in the tenth grade at one suburban public high school in east central Missouri in 1999-2000, 2000-2001, 2001-2002 or 2002-2003. The students included in the study will be those with scores on the Missouri Assessment Program in grades 8 and 10 and TerraNova reading, language, and mathematics in grade 9.

Limitations

The findings in this study will be subject to the following limitations:

1. The study focuses on students from one school district in Missouri. The student sample will be limited to those who meet the criteria for participation.
2. The archival data will reflect the conditions that existed for those cohorts of

students at that time and might not be generalizable to other groups.

3. The sample of students from one suburban high school may not be representative of the population of Missouri high school students in terms of race, gender, urbanicity, socioeconomic status, or course taking.

Significance of the Study

Both increased government spending on education and poor student performance on international assessments have contributed to the accountability system we have in place in our schools today. Policymakers, legislators, educators, parents, and the general public want American students to be the best in the world and to be ready for the demands of postsecondary education and the workplace. Many studies have examined factors, such as socioeconomic status, parents' level of education, and school setting. These are all outside the control of educators. Other researchers have found strong positive relationships between course-taking behavior and student achievement. Even though Missouri students have taken more courses as well as more difficult courses in mathematics, Missouri students' scores on the NAEP and the MAP do not reflect these positive changes in course-taking behavior. Currently, the effectiveness of our public schools in Missouri is measured with test scores on the MAP. It is important to learn all we can at the local level about the alignment of the curriculum and the MAP assessment. Researchers have demonstrated positive relationships between students' opportunity to learn higher order skills and student achievement (Wang, 1999; Wiley & Yoon, 1995).

The MAP mathematics assessment is purported to align with the Missouri Curriculum Frameworks in Mathematics. These Frameworks are based on the NCTM standards. Therefore, the focus in mathematics education in Missouri is to teach a

curriculum that is consistent with the NCTM standards. There is a lack of research that investigates the relationship between the curriculum (as reflected by course taking in mathematics) and performance on standards-based criterion-referenced assessments, such as the MAP. There is also a lack of research that examines individual student's performance on various measures (such as MAP & TerraNova), as well as the relationships between student performance, gender, and course taking. Much of the research on course taking and achievement uses large-scale assessments such as college entrance exams (ACT, Scholastic Aptitude Test (SAT)), or national data sets (NAEP, National Education Longitudinal Study of 1988 (NELS:88), National Longitudinal Surveys (NLS), or High School & Beyond (HSB)).

Much of the gender research examines performance by groups that are not representative of a cross section of students (gifted, college bound); therefore these results are not generalizable to the population. Although there is a significant body of research on gender and mathematics performance, as well as course taking and mathematics performance, there is a lack of research that examines the relationships among course taking, gender, and performance on a criterion referenced test, such as the MAP. Missouri's accountability system for MSIP and NCLB is based on student performance on the MAP at this time. Although researchers have studied correlations between MAP and school finance (Moss, 2003), student background variables (Applegate, 2003), and parent participation variables (Bice, 2002; Laughman, 2000); these are examples of factors outside the control of educators. This research study joins other studies that have examined relationships between student performance on the MAP and student participation in technology classrooms (Bratberg, 2002), standards-based

instruction (LeSage, 2001), and full-day kindergarten (Heavner, 2002). The current study examined the relationship between individual student experience with the curriculum in grades 8 through 10 and student mathematics scores on the MAP mathematics test in grades 8 and 10, taking into account gender. "The NCLB act puts a special emphasis on determining what educational programs and practices have been clearly demonstrated to be effective…"
(http://www.ed.gov/admins/lead/account/nclbreference/reference.pdf, p.11).
In this time of high stakes testing and accountability, educators not only want to know what works, they must know in order to provide all students with opportunities to succeed. This study can provide educators at the local level with a framework for analyzing the relationship between MAP scores and student course-taking behavior.

Summary

This chapter includes the following components: (a) an introduction to the study, (b) background, (c) statement of the problem (d) the purpose of the study, (e) research questions, (f) definition of terms, (g) delimitations of the study, (h) limitations of the study, and (i) the significance of the study. Literature related to mathematics assessment, course-taking behavior, gender studies, and interactions between these factors is reviewed in Chapter II. Chapter III outlines methods that were used in collecting and analyzing the data for this study, information about subjects, research design of the study, instruments, procedures, and human subjects concerns. The results of the statistical analyses will be reported in Chapter IV. Chapter V includes a summary of the study, discussion of the

research findings, conclusions, implications of the findings, and recommendations for future research.

CHAPTER II

Review of the Literature

This chapter examines the literature related to three major themes that are relevant to this study and that will serve as organizational topics. First is the fundamental theme of large-scale assessment as a public accountability measure. This theme is examined in several parts: the historical development of assessment in the United States since 1950; relationships among assessment, curriculum, and instruction; validity and reliability; item types (Multiple choice, Constructed response, Performance events); the use of assessments as accountability measures; and equity and bias. The second theme is course taking. This is examined in six parts including: opportunity to learn (OTL); the relationship of secondary school course taking to performance beyond high school; the mathematics pipeline; ability-grouping and tracking; the year of Algebra completion; and graduation requirements. The third theme is gender, particularly as it relates to performance in mathematics. Gender studies are examined in four parts that include the interaction of gender with the following: assessment, variability of scores, item types, and content strands. Studies that deal with the interaction effects of two or more of these themes will be reviewed in the fourth section. The chapter will close with a summary of the research and an outline of chapters three, four, and five.

Assessment

This section of the review examines literature on the effects of using standards based, state-mandated assessments for accountability. A historical view of testing from 1950 to 2004 in the United States is presented. The development of standards and related assessments is discussed as well as the validity of score-based inferences. High-stakes

tests have both intended and unintended consequences for curriculum, instruction, and classroom assessment. The results of the literature review are related to the Missouri Assessment Program (MAP), the criterion variable used in this study. The MAP is a state-mandated program that has high stakes for school districts. The MAP includes three types of items: multiple choice (MC), constructed response (CR), and performance events (PE).

Students in Missouri are required to take the MAP mathematics test in grades 4, 8, and 10. The students' scores on these assessments are a component of the Performance Points Matrix of the Missouri School Improvement Program (MSIP), the system by which Missouri's public schools are accredited. Although Missouri does not attach significant student consequences for individual test performance, some local districts offer incentives to students to encourage high scores. Students' test scores are used to determine the instructional effectiveness of schools and districts. This qualifies MAP as a high-stakes test.

Historical Development of Testing in the United States

Robert L. Linn (2000) traced the development of testing in America and found that there has been a different wave of testing in the United States for each decade in the last half of the twentieth century. The following is a summary of his research and his interpretation. In the post World War II era of the 1950s, testing provided ways to track students into either a vocational or a professional program within a comprehensive high school. In 1965, as part of President Lyndon Johnson's "Great Society," the 89th Congress of the United States passed the Elementary and Secondary Education Act (ESEA). The

main intent of ESEA was to provide funding for programs that would allow equal access to a quality education for all elementary and secondary students in America, especially those students who were most disadvantaged. Most of the $1 billion per year allocation from ESEA went to Title I, which funded programs to meet the needs of children from low-income families. This financial assistance was distributed to 90% of all schools in the country, including non-public schools. Senator Robert Kennedy was among legislators who wanted a means for evaluating the Title I programs in order to hold schools accountable (Carleton, 2002). The Title I Evaluation and Reporting System (TIERS) was developed for this purpose. The use of the Normal Curve Equivalent (NCE) scale for reporting test scores became prevalent with the TIERS program (Linn, 2000). Title I students were often tested in the fall and the spring of the same school year in order to measure academic growth. This growth was then attributed to the Title I program. Teachers were very interested in seeing their students do well on the spring assessment so that money from the government would continue to fund the program. Linn (2000) describes TIERS as an early version of a high-stakes test that led to corruption of indicators and loss of valid results because of the teacher practice of teaching to the test. While government funding for education was on the rise, increased public attention was being focused on ways to measure the effectiveness of our educational system.

In the 1970s and early 1980s minimum-competency testing (MCT) became popular. As the name suggests, the skills tested were at very low levels of achievement. By 1983, thirty-four states had some form of MCT. Florida's program was very controversial because of the differential passing rates for African American, Hispanic,

and White students. This spawned discussion about opportunity to learn issues that still exist today in the differential results of tests that are currently in use.

A norm-referenced test is also commonly referred to as a standardized test. During the development of such tests, the test is administered to a sample group and the distribution of the sample group's scores is compared statistically to what is called a normal distribution, one where the scores fall in a bell shape where the mean, median, and mode are equal. Test items are considered 'good' items if they discriminate well and allow the scores to fall on the desired normal or bell curve. This process is sometimes known as "norming" the test. New test-takers scores are then referenced to the norm group and reported as they relate to that normal distribution (Berlak et al., 1992).

Linn (2000) reported that performance on norm-referenced high-stakes tests typically improves for the first two to three years and then levels off unless the test is renormed or a new form of the test is published. In that event, there is a sharp decline in scores followed by the same trend seen on the previous cycle. This pattern in norm-referenced scores is a result of the instructional practice of teaching to the test, which inflates scores.

In Missouri, the MC questions on the MAP come from the Survey portion of the TerraNova, a norm-referenced standardized test, published by CTB McGraw-Hill. The Missouri Department of Elementary and Secondary Education (DESE) issued a news release in 2001 that stated that the national average TerraNova score was the 50^{th} percentile and then gave the percentiles for Missouri students at each grade level tested on the MAP. In mathematics, the data at the state level from 1998-2001 showed the median percentiles for fourth graders going from 56 in 1998 to 62 in 2001. Similarly,

grade 8 went from 56 to 60 and grade 10 went from 66 to 70 (http://dese.mo.gov /news/2001/terranova.htm). The TerraNova has not been renormed since Missouri began using it as part of the MAP tests.

In the late 1980s and through the 1990s norm-referenced standardized tests became the primary accountability measures for schools. With the pressure of accountability, the teachers and administrators did all they could to increase student achievement scores. The upward trend continued until educators were faced with what has been named the Lake Wobegon effect, where all students are "above average." In almost every state and most school districts, most of the children were scoring above the national norm! While teachers may not have given their students the exact information that was on the test, many narrowed the curriculum to only what was sampled on the test. Most of what was on these tests was still at a basic skill level; therefore the curriculum was becoming less rigorous at the same time it was being narrowed (Mehrens & Kaminski, 1989; Popham, 2001).

The widespread practice of teaching curriculum that was narrow and shallow continued throughout the 1970s and 1980s and led to the standards-based accountability systems schools have today. Researchers and test developers as well as many parents and legislators felt that performance assessments would be more suitable for assessing higher-level skills and would lead to a different kind of instruction. (Berlak et al., 1992; Mehrens, 1992; Wiley & Yoon, 1995). The thinking was (and in some cases still is) that if the teachers would teach to the skill levels assessed on performance tests, then they would necessarily teach the higher order thinking skills. The theory is that these instructional practices would in turn benefit the students by helping them acquire higher-

level skills.

Educators and test developers share concerns that tests with only multiple choice items might not effectively measure what students know and are able to do (Darling-Hammond, 1985; Lukhele, Thissen, & Wainer, 1994; Pearson & Garavaglia, 2003). On the other hand, there are concerns about the costs in time and money to administer and score tests with open-ended items, such as constructed response and performance events. In one analysis of Advanced Placement (AP) Chemistry and AP History tests, researchers found that "a constructed response test of equivalent reliability to a multiple-choice test takes from 4 to 40 times as long to administer and is typically hundreds to thousands of times more expensive to score" (Lukhele et al., 1994, p. 234).

Federal legislation has mandated accountability of educators to the public they serve. The federal government has charged state departments of education with developing standards and assessments that will serve their state's districts, schools, and students while meeting the accountability demands of the federal laws. Changes in the types of statewide assessments used for accountability have led to changes in curriculum and instruction as well as changes in classroom assessments.

Relationships among assessment, curriculum, and instruction

The structure of American education can be viewed as one that includes tested, written, and taught curriculum. Educators strive to keep these three components in balance so that none of the three dominates the other two. This section will include a discussion of research related to efforts at aligning assessment, curriculum, and instruction.

The development of assessments through the last half of the 20th century led to standard-based performance tests. In Missouri, the Outstanding Schools Act of 1993 required the development of state standards, the Show-Me Standards, and an assessment that would measure student mastery of those standards, the Missouri Assessment Program (MAP). School districts are charged with the responsibility of developing curricula that incorporate these standards for all students, regardless of their course-taking behavior. Classroom teachers are responsible for delivering instruction that meets all of the following requirements: the instruction should lead to student mastery of the course objectives, prepare students for their future, be consistent with the standards, and prepare students to be successful on the MAP test.

Researchers have examined the processes used to develop standards and found the language used in the standards to be a frequent source of teachers' frustration and confusion in interpreting and implementing the standards (Fan & Chen, 1997; Hill, 2001; Linn, 2000; Moss & Schutz, 2001; Rothman et al., 2002). For example, in one study, the language of the standards was not clearly understood by a group of elementary math teachers, causing them difficulty in writing curriculum, planning instruction, and developing local assessments to support the standards (Hill, 2001).

The process for developing standards varies from state to state. A committee of administrators and content specialists at the state level usually writes the standards. Often teachers are included on the committee. The process involves group discussion, which leads to a consensus about the wording of the standards. In an examination of this practice, researchers contended that consensus masked diversity. They felt that a consensus-seeking discourse did not reflect the standards development process and that

dissensus would be valuable for equity and balance. These researchers recommended that some of the dissenting views should be used as examples or non-examples to help clarify the standards for the reader (Moss & Schutz, 2001). In Missouri, a draft of the standards was sent to educators to solicit their input and feedback. The final document was modified from the original draft but did not include clarification in the form of examples and non-examples as recommended by Moss and Schutz (2001).

Missouri's Outstanding Schools Act of 1993 required the development of the state standards. The Missouri Show-Me Standards consist of 73 standards: 33 performance (process) standards and 40 knowledge (content) standards. Six of the content standards are in mathematics. The No Child Left Behind Act of 2001 ("NCLB," 2002) mandated that states develop frameworks and assessments and submit their NCLB plans. Effective June 10, 2003 all fifty states, the District of Columbia and Puerto Rico had NCLB plans that were approved by the U.S. Department of Education. The Missouri NCLB plan uses the MAP tests in Mathematics and Communication Arts to monitor student mastery of the Show-Me Standards (http://www.ed.gov/new/pressreleases/2003/06/06/02003.html).

The MAP mathematics assessments contain three types of items that assess the six content strands in mathematics. The students may receive 0 or 1 point on a multiple-choice question; 0, 1, or 2 points for a constructed response question; and 0, 1, 2, 3, or 4 points on a performance event. These are raw score points. The information about the weighting of individual questions that leads to the scale score is not made available. The table on the next page shows the distribution of raw score points by content strand and item type for the MAP mathematics test in grades 8 and 10 in the spring of 2003.

Table 3

Spring 2003 MAP Mathematics
Percent of total raw score points for each Content Strand and Item Type

Content Strands	Grade 8 MAP	Grade 10 MAP
Number Sense	25%	22%
Geometry & Spatial Sense	21%	21%
Data Analysis and Probability	19%	16%
Patterns & Relationships	13%	16%
Mathematical Systems	9%	14%
Discrete Mathematics	12%	11%
Item types	**Grade 8 MAP**	**Grade 10 MAP**
Multiple Choice	41%	34%
Constructed Response	48%	55%
Performance Event	11%	11%

A recurring theme in the research is that what is not tested is not taught (Darling-Hammond, 1985; Hoover, 1998). Instructional issues surrounding assessment often relate to the narrowing of the curriculum and the extensive use of time to prepare students for taking the state test (Gordon & Reese, 1997). Teachers report frustration about spending so much time preparing students for a test and sometimes being forced to teach only what is on the state test. Many researchers have expressed concern about assessment dominating instruction and upsetting the appropriate balance between the written, taught and tested curriculum (Darling-Hammond, 1985; Kulik, Kulik, & Bangert, 1984; Madaus, 1994; Popham, 2001). One study concluded, "high-stakes testing has become the object rather than the measure of teaching and learning, with negative side-effects on curriculum, teacher decision making, instruction, student learning, school climate, and teacher and student self-concept and motivation" (Gordon & Reese, 1997, p. 366).

Popham (2001) especially dislikes what he calls "instructionally corrupt test

preparation" (p. 23). This occurs when teachers either design their instruction around actual test items…or teach toward *clone* items. However Gordon & Reese (1997) found, as many other researchers have, that even when students can be taught to answer test questions correctly they may still not have learned the important content behind the answer (Fuchs, Fuchs, Karns, Hamlett, & Katzaroff, 1999; Linn, 2000; Rothman et al., 2002). Teaching to specific test items is not possible with the CR or the PE portions of the MAP test because those two portions of the tests may differ from year to year and multiple forms of a test may be administered in a given year.

Other researchers (English, 2000; Hoover, 1998) maintained the importance of using item analysis data in curriculum design and instruction. This level of analysis is not possible in Missouri because the data educators and parents receive for MAP does not include specific test items and individual student responses. Although the number of points an individual student received on a specific MAP item is available on the reports at the district and building level, the test items themselves are rarely released.

When analyzing mathematics instruction and performance assessment, it is important to consider the mathematics education reform movement and what is known as standards-based instruction. The standards that were used in developing the Curriculum Frameworks for Mathematics in Missouri are the National Council of Teachers of Mathematics (NCTM) Standards. Although NCTM (2000) released an updated version of the standards, the information was rearranged but the message was much the same as that of the 1989 standards (NCTM, 1989). The NCTM message is related to instructional practices that promote teaching for student understanding of mathematical concepts, rather than defining a list of what objectives to teach.

Two studies examined the relationship between student achievement and standards-based instruction. Elementary students whose teachers reported delivering standards-based instruction performed better on the MAP than students whose teachers did not report that type of instruction. In this study of MAP mathematics achievement by a sample of fourth grade students in Missouri, the highest correlations between student achievement and a specific component of standards-based instruction involved students having "opportunities to clarify and justify their ideas through oral or written dialogue" (Le Sage, 2001, p. 93). In the second study, student achievement was measured using traditional assessments. Mayer (1998) compared the effects of teaching in a manner consistent with the NCTM standards to teaching in what the author called a "traditional classroom." The students were middle school and high school Algebra students. Middle school students with NCTM standards-type teachers showed the most growth, with high-ability middle school students benefiting the most. The high school students were neither helped nor hindered by NCTM standards-type instruction.

Sixteen elementary school teachers participated in an experimental study that sought to determine if classroom instruction driven by the use of performance assessments (PA) had an effect on students' problem solving in mathematics. The findings showed that in the experimental group above grade-level students showed improved skills in three assessed measures of problem solving, at grade-level students on two and below grade-level students on only one. The results for below grade-level students were not significantly different in the PA group than in the non-PA group. There was a recommendation for more staff development that would be specifically aimed at reaching low achieving students (Fuchs et al., 1999).

Research indicates that many teachers who think they are delivering instruction in a way that will allow students to construct their own meaning may be fooling themselves. A constructivist approach and a reliance on metacognition are in much of the research (Resnick, 1987; Young, 1997). There is also research that indicates that teachers sometimes feel that if they have cooperative groups and use manipulatives, then they are teaching in a standards-based way or in a new way (Cuban, 1984; Cohen, 1990).

The case study of Mrs. Oublier (Cohen, 1990) pointed out many paradoxes in the reform movement. Mrs. Oublier, her principal, and her assistant principal all felt she was using the new teaching methods but she really only had students handling beads and sitting in groups as she delivered very direct instruction. In their evaluation, the administrators noted no mismatch between the desired student-centered instructional practices and a teacher evaluation model that was built around a direct instruction method of teaching that was very teacher-centered.

In her study, Lee (1998) found that an analysis of eighth-grade mathematics practices in two states showed that linking student assessment and texts to the state frameworks is positively correlated with the level of progressive instructional practices in mathematics classrooms. In Salmon's (1997) study of six teachers, she found that the teachers focused on how to teach and assess and that the state standards and assessment determined the curriculum they taught. She found that only two of the six teachers in the study actually gave students the types of problems that demanded higher order thinking skills. This occurred despite the fact that the teachers indicated their understanding that performance assessments are important because they are ill-structured tasks that cause students to think analytically and demonstrate proficiency as in real-life situations.

Visintainer (2002) studied the relationship between elementary students' performance on both a criterion-referenced state-mandated test, the Maryland School Performance Assessment Program (MSPAP), and a norm-referenced test, the TerraNova. She found that,

> to the extent that MSPAP has actually 'driven' instruction in Maryland's elementary classrooms for nearly ten years, it seems, at worst, to do no harm. If Maryland's teachers are 'teaching to the test' for MSPAP, what is being learned is not incompatible with that tested by the nationally-normed (and marketed) TerraNova. (Visintainer, 2002, p. 69)

Hoover (1998) felt that the new assessment in Pennsylvania was shaping instruction. At the same time, the teachers who responded to his survey had reservations about the instructional effectiveness of classroom performance assessments. Several studies joined his in showing that the stability of teachers' beliefs seemed to be a strong determinant of the degree to which instructional practices might change (Fairman, 1999; Levine, 1998; Salmon, 1997; Hoover, 1998).

Hoover (1998) also made a strong connection between tested curriculum and taught curriculum. He stated emphatically that what is tested is what we value and that test results can and do change reality in schools. He maintained that there is nothing wrong with teaching to the test if the test matches the objectives in the curriculum. English (2000) went further by stating that if the tests are to be considered valid, then we must teach to the tests. Researchers agree that the curriculum must be aligned to the state standards and assessments. State and federal laws mandate such an alignment as well. The type of alignment of curriculum and instruction to the Show-Me Standards and the

MAP in Missouri does not take into account the differential course-taking behavior of students. There is one set of standards, one state assessment, and many different course-taking opportunities for students between and within schools.

Validity and reliability

The advent of new assessment types has spawned controversy over the reliability and validity of performance-based assessments. In addition to concern about costs in time and money to administer these assessments, there are concerns about the setting of achievement levels and cut scores as well as the consistency of scoring (Pomplun & Sundbye, 1999). The policy assumption in Missouri is that the MAP is strongly aligned to the content in the Show-Me Standards. Missouri has not provided an alignment of the state's mathematics content standards with the competencies assessed on other instruments, such as the ACT or the SAT. There are no published studies that examine the external validity of the MAP.

The Missouri Department of Elementary and Secondary Education (DESE) and CTB McGraw-Hill used the "bookmark procedure" to set the five achievement levels for the MAP tests.

> A panel composed of 40 to 45 teachers, parents, and business professionals reviewed the rank ordered test items from field-testing of the MAP. Test items were rank ordered from easiest to the most difficult based upon student performance during the field test. The panelists placed a bookmark at the point that they thought a student performing at

> Advanced, Proficient, Nearing Proficient, or Progressing would perform. The panelists then discussed the rationale for their judgments. The judgments of the panel members were averaged to establish cut off points for each achievement level. (Bratberg, 2002, p. 11)

A staff member at DESE provided information on reliability and validity of the MAP tests (W. Gerling, personal communication, July 6, 2003) which is available on the web at the following address: www.dese.state.mo.us/divimprove/fedprog/discretionarygrants/ReadingFirst/DMAP.pdf There were reliability coefficients reported for the MAP assessments in each content area at each grade level. An appendix included reliabilities for standardized measures such as the ACT, the SAT, the SAT-9, and AP exams in several areas. Neither AP Calculus nor AP Statistics was included. The MAP mathematics scale score reliability coefficients for grades 8 and 10 ranged from 0.927 to 0.931 for grade 8 and 0.929 to 0.940 for grade 10. The ACT mathematics had reliability coefficients from 0.89 to 0.91. Tests that included open-ended items showed lower reliability for those parts of the test. The SAT-9 had overall reliability coefficients from the middle 0.80s to the 0.90s; however, the open-ended items for that test had lower reliability coefficients ranging from the 0.60s to the low 0.80s. There were similar differences in reliability coefficients for open-ended items and composite scores for AP Government, AP History, and AP English. The document contained statements about CTB McGraw-Hill and DESE conducting validity studies on the MAP to ensure that the items measured the constructs they were intended to measure. However, there were no data included to support the claims of validity of the MAP tests.

Kane (1994) and Popham (2001) both assert that the question of validity does not

relate as much to test items as it does to the interpretations and inferences we make from the results. In terms of high-stakes tests and performance standards,

> The passing score is a number and the performance standard is a construct…. The aim of the validation effort is to provide convincing evidence that the passing score does represent the intended performance standard and that this performance standard is appropriate, given the goals of the decision process. (Kane, 1994, p. 433)

There is no "passing score" on the MAP, but the state has defined Proficient as "the desired achievement level for all students. Students demonstrate the knowledge and skills called for by the Show-Me Standards." The following is the statement DESE makes in describing the Advanced level: " Students demonstrate in-depth understanding of all concepts and apply that knowledge in complex ways" (http://dese.mo.gov/schooldata/). The requirement of NCLB is that all students' scores in mathematics be at the Proficient level or above by 2014. Missouri has chosen, for now, to continue to use the same achievement levels for MAP mathematics that they have used since its inception. Proficient or above could be interpreted as being a "passing score" for accountability purposes.

Popham outlines procedures for internal and external validity checks. Internal validity focuses on the consistency of the results and speaks to the descriptive assumption. External validity compares the results with some external measure of competence. This is not often used because it is often difficult to find external objective measures of competency. External checks on validity address the policy assumption.

Kane asserts that these are mostly reality checks that can find major flaws but "would not be sensitive to small shifts in the passing score" (Kane, 1994, p. 457).

The Lukhele et al. (1994) study of the AP scores found that the CR items only yielded a small amount of information at the high end of the scores but decisions were made at the 2-3 point range where the cut score is. "This observation suggests that the test is somewhat misaimed and is too difficult for the decision task for which it was built" (p. 248). A study was done to examine the external validity of the reading test that is part of the Kentucky Instructional Results Information System (KIRIS) by comparing KIRIS scores to the ACT reading scores for a group of 2,668 twelfth-grade students (Strong & Sexton, 1996). The KIRIS identified only 8.54% at the top two levels (Proficient and Distinguished) but 29% of the students in the sample scored between 24 and 36 on the ACT. These ACT scores would fall in the top 25% of the nation. The researchers concluded that this was a failure of the KIRIS to discriminate at the high end of the distribution. They added that similar failure was noted at the low end of the score distribution. While Kentucky's test may have effectively assessed mastery of Kentucky's standards, it did not seem that those standards were well aligned with ACT objectives.

The only external validity information available to date on the MAP mathematics is related to Missouri students' performance on the National Assessment of Educational Progress (NAEP). In 2001, 21% of Missouri eighth-graders scored "Proficient" on the NAEP and 28% scored Proficient in 2003. In response to that information, the Missouri Department of Elementary and Secondary Education (DESE) issued a news release, dated

November 13 2003, in which they quoted Missouri's Commissioner of Education, Dr. King:

> The NAEP assessment is challenging for kids. It has a structure and expectations that are similar to Missouri's MAP tests. The proficiency scores on both exams are similar, so we believe the NAEP scores offer an important verification that we are 'on track' with our state testing standards. (http://dese.mo.gov/news/2003/naepscores.htm)

Item types

This section will review the research related to item types. The MAP contains Multiple Choice (MC), Constructed Response (CR), and Performance Events (PE). Researchers are divided about the relative merits of various item types. Most agree that MC items do not offer students an opportunity to demonstrate what they know and are able to do (Darling-Hammond, 1985; Lukhele et al., 1994; Mehrens, 1992; Pearson & Garavaglia, 2003). Many also found that CR and PE items are not cost effective and may not provide enough additional information to warrant the cost in time and money to administer and score them (Linn, 1993; Lukhele et al., 1994; Mehrens, 1992).

In an early discussion of the use of performance assessment for accountability purposes, Mehrens (1992) named factors that led to support for performance assessment. Most of these factors center on dissatisfaction with the multiple-choice format. Some of the concerns were related to the negative effects of teaching to the multiple-choice tests and concerns about delimiting domains being assessed when the questions are in the multiple choice format. In his discussion about these issues, Mehrens states that multiple-

choice assessments are able to effectively assess knowledge that is a necessary, but perhaps insufficient, condition for acquiring expertise. Cognitive psychologists promote performance assessments to measure procedural knowledge. However, Mehrens cautions that these types of assessments will test narrower domains because it will be necessary to include only a few of these types of items in one assessment because of the high cost in student time to take these tests and professional time to score them. Two of the many reasons cited in favor of the use of performance items in addition to multiple choice formats are: The Lake Wobegon effect ("raising scores without raising the inferred achievement" (Mehrens, 1992, p. 4)) and educators' beliefs that teaching to a performance test would lead to beneficial instruction of procedural knowledge, higher level skills, and critical thinking.

Linn (1993) describes performance assessment items as very task specific. He illustrates how efforts to increase generalizability by increasing the number of raters show no gains. Increasing the number of topics or the number of tasks showed great gains in generalizability. However, increasing the number of tasks may be cost prohibitive. Missouri's MAP includes two performance events at each level each year. Few of these items are released because of the high cost of developing new items.

Researchers also sometimes question whether items measure the construct they purport to measure. An example would be "mathematics items in which the reading is more a factor than the mathematics" (Rothman et al., 2002, p. 15). The intent of having students write on a mathematics assessment may be to encourage students to communicate effectively. If an item puts more emphasis on language than mathematics it is not testing mathematics achievement but language achievement.

Assessments as accountability measures

Gordon and Reese (1997) "define high stakes tests as standardized achievement tests used as direct measures of accountability for students, educators, schools, or school districts, with significant sanctions or rewards attached to test results" (p. 345). Bishop (1996) suggested that the United States is out of step with most other advanced countries, where the curriculum is assessed by examinations that are graded at the national or regional level. He claimed American students are judged by internal standards such as class rank or grades. He recommended, "statewide assessment of competency and knowledge that are keyed to the state's core curriculum should be made a graduation requirement" (pp. 104-105). Missouri has adopted a statewide assessment of standards but it has not adopted an official state curriculum nor has it tied MAP to graduation requirements.

Levine studied the use of performance assessments as tools for reform in an urban school district. Levine (1998) suggested that the best system of accountability is a multi-layer system of two-way accountabilities. The layers are student-teacher, teacher-principal, and principal-central administration. Although he felt that in his model all students can learn and teachers could deliver effective instruction, it would not happen without the support of required resources of time and materials. He recommended that these resources be demanded as part of the two-way accountability system. His message clearly was 'do not demand results if you do not deliver on resources.' The layers of accountability could be extended to the local board of education, taxpayers, the state department of education, or the U.S. Department of Education. In 2003, the Missouri Department of Elementary and Secondary Education (DESE) stopped requiring schools

to administer the MAP in Social Studies and Science, but allowed schools the option of giving the tests. The state no longer had the resources to continue to fund the MAP tests in those areas. Science will be funded again by DESE when it becomes part of the NCLB mandate in the spring of 2008.

In a small qualitative study of six elementary teachers, Salmon (1997) found the purpose of performance assessments was to make schools accountable for helping students acquire higher order thinking skills, such as analysis and synthesis. On the other hand, in his recent book, Popham (2001) contends that talk about using tests to drive instruction is rhetoric. He feels the real rationale for state-mandated high-stakes tests is the accountability of school districts to state departments. These tests are the yardsticks being used to judge teachers and districts and he contends no one cares very much how students are instructed as long as there are positive results on the state assessments. In interviews with principals, Hoover (1998) found that principals' perceptions of students' accountability was much higher than teachers' perceptions of student accountability.

Equity and Bias

Test developers strive to write assessments that will be free of cultural and gender bias. This section will report data and review studies that examined differential performance by socioeconomic status, gender, or ethnicity. In a qualitative study with pre-school children in the United Kingdom, Cooper (1998) found that children's ways of answering open-ended questions related to their social class differences. These results contradicted the notion of many test-developers that all cultural bias in tests can be eliminated.

Although Hoover (1998) contended that performance assessments promote educational equity, that claim was not substantiated by his study. He asserted that comparison of school outcomes should be considered carefully. He pointed out that different schools have different populations, different resources, and different educational goals.

Evidence of the relationship between these 'differences' and MAP scores can be found in the school and district data reported by DESE. Missouri school districts can be recognized annually as having "distinction in performance." In 2003, 176 districts were so recognized, 12 of those were in the same county as the district participating in this study.

> The 176 districts will receive the "Distinction in Performance" award, based on criteria set by the State Board of Education. The annual recognition is based on school districts' performance on MAP test scores, ACT test scores, attendance and dropout rates, and other measures of academic performance during the past school year (2002-03).
>
> To qualify for the recognition this year, K-8 districts had to meet 5 out of 6 performance standards (at least 45 out of 54 possible points), including all of the standards that are based on MAP test scores. K-12 districts had to meet 11 of 12 standards (at least 91 out of 100 possible points), including all of the MAP-based performance measures.
>
> http://go.missouri.gov/press/press121803f.htm

The district participating in this study is one of 23 suburban districts in the county. When these 23 districts are ranked in order of the percent of students who receive

free or reduced lunch (FRL), those districts with "distinction in performance" rank number one (lowest percentage of FRL) to 11 and one district ranks 18th. However, the district that ranks 18th also ranks fifth highest in per pupil expenditure.

Conversely, Missouri school buildings can be designated as "academically deficient" if the performance for two consecutive years places the school in the lowest 50 schools when considering the percent of students who score in the bottom two levels on the MAP tests. Only three schools received that designation in 2003 and all are above the state median for the percent of students who received free or reduced lunch. These three schools also serve 86% to 95% non-Asian minority students. Although the MAP is not intended to be a measure of socioeconomic status, for the most part school scores fall in line with the percent of students receiving free or reduced lunch.

The intent of federal and state legislation that led to state systems of accountability was to provide a quality education to all students. The data for the school with the highest MAP 10-mathematics achievement in the state points to possible bias. This school is an urban magnet school serving a population that is described as highly motivated and college bound. The admission criteria are selective and based on ability, achievement, as well as residency and race. The district reported that 52 of 54 graduates in 2003 scored above the national average on the ACT, there were no dropouts, and 96.3% of the 2003 graduates enrolled in college. High graduation rates, a high percentage of enrollment in post-secondary education, and a high percentage of students scoring above the national average on the ACT are all acknowledged markers of a highly successful student population.

In 2003, this school had 50 students accountable for grade 10 MAP mathematics scores. The racial makeup of the class was 6% Asian, 48% White and 46% Black. While only 2.5% of the students in the district scored in the top two levels, 54% of the students in this school scored Proficient or Advanced on the grade 10 MAP mathematics test, the highest percentage in the state. As a magnet school, this school has a population that is not representative of the entire district in ability or background variables. For instance, while the district reports 83.21% Free or Reduced Lunch (FRL), the school has only 18.26% FRL. Also, the district reports 81.1% Black students, while the 10th grade in this building has 46% Black students. Even with overall outstanding external measures such as 0% dropout and high ACT scores, and with the highest percentage of students in the state at the top two levels on the MAP mathematics test, the Black students at this school are not equally represented in the top two levels. Only 34.8% of these Black students scored Proficient or Advanced, while 70.8% of the Whites, and 66.7% of the Asians achieved those high levels of performance. It appears that while the Black students are performing well on these external measures, their performance on the MAP is not consistent with their ACT scores, graduation rates, and college attendance.

Course taking

Research related to student course-taking behavior will be examined in six parts. Some studies related to course taking will overlap with the issues of assessment and gender. The studies reviewed in this section will be organized under these topics: opportunity to learn (OTL); the relationship of secondary school course taking to

performance beyond high school; the mathematics pipeline; ability-grouping and tracking; the year of Algebra completion; and graduation requirements.

Opportunity to learn

Students cannot learn subject matter if they are not enrolled in the appropriate courses where that subject matter is taught. Opportunity to learn (OTL) is influenced by course offering as well as course taking. Researchers have found that course offerings are not equitable for all students and are not consistent at all schools. Socioeconomic status, school size, and urbanicity sometimes contribute to availability of courses (Oakes et al., 1990). Research related to course offering and OTL will be reviewed in this section.

> Course taking is the most powerful factor affecting students' achievement that is under the school's control…although schools cannot do much about the social class of the students who attend them, they can do something about the patterning of courses and the procedures used to place students in classes. . . schools can influence the achievement of students, even when the social-class origins of the students they serve may not be conducive to achievement, by restructuring the patterning of classes and facilitating the placement of students in more challenging courses. (Spade, Columba, & Vanfossen, 1997, p. 125)

Oakes et al. (1990) stated that our nation rejects the notion that we should provide less to those who are less advantaged or less able. The recent NCLB legislation affirms that position by mandating that 100% of American students be proficient in reading and mathematics by 2014. These researchers examined statistics on course offerings and course taking in mathematics and science, as well as

statistics on who teaches at each level. They found Algebra in junior high and Calculus in high school to be critical gatekeeping courses. They studied interactions between race, socioeconomic status (SES), and tracking on OTL and concluded that the

> Quality of learning opportunities available to different categories of children related strongly to the social and economic circumstances of children's' families and communities. (Oakes et al.,1990, p. iv)

Wang (1999) conducted a study of longitudinal California Test of Basic Skills (CTBS) data on 2,443 eighth-grade students in a large urban district in California. There were two objectives: to determine how Limited English Proficiency (LEP) and immigrant status affected course taking and to determine how course taking interacted with language proficiency and immigrant status to affect mathematics achievement. Some important results of this study were:

1) When course taking was equalized, girls' mathematics performance was statistically lower than boys.
2) Although SES differences accounted for some variation in scores, they were less important than other student characteristics.
3) Students' course taking explains mathematics achievement even after considering students' descriptive characteristics, language proficiency, immigrant status, and SES. Students who studied Algebra, honors mathematics, or elective mathematics had significantly higher test scores than students enrolled in standard mathematics classes. Students who took a minimum standards course performed significantly lower than students with a

> standard mathematics class. Students with elective mathematics (doubling up) had the highest growth rate; students in minimum standards courses had slowest growth rates.

In addition, Wang (1999) found that "students who entered the sixth grade below average in mathematics achievement were likely to fall further behind students entering the sixth grade with above-average mathematics achievement by the end of eighth grade" (p. 44). Growth rates for various groups ranged from 11 points per year to 19 points per year. Wang refers to this phenomenon as 'fanning' (Wang, 1999, p. 43).

The amount and intensity of course offerings in schools directly affects students' course-taking pattern and is therefore a part of the OTL research. Several researchers have found that a constrained curriculum where there are fewer course-taking options leads to higher percentages of students taking more rigorous courses (Ayalon, 2002; Finn, Gerber, & Wang, 2002; Lee & Bryk, 1988).

Not surprisingly, researchers have consistently found that students who take more math courses and more rigorous math courses score higher on measures of math achievement and show greater growth over time as they go through grades 9 through 12 (Jones et al., 1986; Rock & Pollack, 1995a, 1995b).

> There appears to be consensus among researchers that quantity of schooling is positively related to academic achievement. Whether achievement is measured by ACT, SAT, or tests developed for NELS and HSB, higher test scores are associated with spending more time in related course work. (Goertz, 1989, p. 7)

Wise (1985) conducted a study using data from *The Project TALENT Women and Mathematics Study*. She found that 9th grade math achievement was the strongest predictor of twelfth-grade math achievement (r = .78) with math courses as the second strongest predictor (r = .73). When these two factors were combined, the multiple correlation of math courses taken and ninth grade achievement with twelfth-grade achievement was .84, accounting for just over 70% of the variation.

Relationship of secondary school course taking to performance beyond high school

Educators recognize high school mathematics courses as important stepping-stones to success in college and employment. Research indicates that students' high school course taking in mathematics prepares them for success in college-level mathematics courses (Adelman, 1999; Long, 2003; Rose, 2001; Roth et al., 2001; Schiller & Muller, 2003). In fact, Roth et al. (2001) found that taking more mathematics in high school, even if it means lower GPA, led to higher scores on the junior college mathematics placement test in their Florida study.

Other researchers (Pelavin & Kane, 1990) have defined enrollment in Geometry in high school as a strong correlate of college enrollment and completion. They studied HSB data and found that "83% of the students who took Geometry matriculated" (p. 75). In examining course taking for different racial groups, they found that 80% of black students who took Geometry attended college within four years of graduation and the rate was 82% for Hispanic students. They stated, "the gap between minorities and whites virtually disappears among students who took geometry" (p. 76). They also found that in a sample of 15,941 students studied, only 5% with less than one year of Geometry

"attained a Bachelor's degree or senior status within four years of high school graduation" (p. 78). Those percentages were lower for minority students: only 2.5% of Black students and less than 2% of Hispanic students without a Geometry course finished college or attained senior status within four years of high school graduation. Adelman sharply criticized Pelavin and Kane's analysis because it was based on incomplete history (3.5 years after high school graduation) and what Adelman called "a dependent variable that is far from the desired end of the story [senior status or a bachelor's degree]" (Adelman, 1999, I. Cultivating ACRES: The academic resources index Section, HIGHMATH: Getting beyond Algebra II subsection). Adelman went on to say that the 29% rate of Geometry students who earned a bachelor's degree was a constant rate, even seven years later, and the percentage (29%) was well below the rate for students who completed levels of math higher than Geometry in high school.

Adelman merged HSB twelfth-grade test scores, high school class rank, and academic curriculum intensity (itself a complex variable) and created a variable called "academic resources" or ACRES. In reporting each of the three component variables in quintiles, he demonstrated that the highest mathematics course taken in high school was the most powerful predictor of bachelor degree completion, followed by twelfth-grade test scores and then class rank. When he added in a socioeconomic (SES) factor, it edged out class rank for third place as a predictor variable. The long-term bachelor's degree completion rate (by age 30) for ACRES was 72.5% for the highest quintile versus 55.5% of the highest quintile of SES. Students from the lowest two SES quintiles who are in the highest ACRES quintile earn degrees at a higher rate (66% and 62.2%) than the majority of students in the highest SES quintile. Those in the top SES quintile but in the 3rd, 4th,

and 5th quintile for ACRES earned degrees at 51.2%, 28.1% and 12.8% respectively. Low ACRES students earned degrees at a low rate regardless of their SES quintile.

> Of all pre-college curricula, the highest level of mathematics one studies in secondary school has the strongest continuing influence on bachelor's degree completion. Finishing a course beyond the level of Algebra 2 (for example, Trigonometry or Pre-calculus) more than doubles the odds that a student who enters post-secondary education will complete a bachelor's degree. (Adelman,1999, Executive Summary Section, Selected Findings subsection, ¶3)

Researchers conducting studies of student success in college mathematics at two-year colleges found that the highest course taken in high school was more important than scores on placement tests or degree intentions of students (Berry, 2003; Long, 2003). One researcher recommended that "High schools should find an alternative to tracking, less rigorous mathematics courses…do not prepare students for anything except to receive a high school diploma" (Berry, 2003, p. 406). Long (2003) found that students who placed into courses lower than College Algebra at the community college had less than a 5% passing rate when they took College Algebra as a subsequent course.

Employers are concerned about the need for job candidates, who are proficient in mathematics to fill positions in the workplace. In a "white paper" called *Mathematics Equals Opportunity,* the results of a survey were reported stating that students who were given the Armed Services Vocational Aptitude Battery (ASVAB) and scored in the top quartile were less likely to be unemployed and likely to earn more, even if they did not pursue post-secondary education. There was no direct link made between those scores

and course taking in mathematics. In addition, the white paper stated that manufacturing businesses were calling for entry-level automobile workers "to be able to apply formulas from algebra and physics to properly wire the electrical circuits of any car" (p. 15). The white paper also stated "computer technology and health services are fields that can require substantial mathematics and science preparation" (U.S. Department of Education, 1997, p. 15). The Bureau of Labor Statistics reports the 'hottest jobs' on its web site. Of the top ten fastest growing occupations, three are computer related and six are in the health care field (http://www.bls.gov/emp/emptab3.htm). Preparation for most jobs in technology and health care requires, at a minimum, an understanding of algebra.

Mathematics Pipeline

The mathematics pipeline is a metaphor for the accelerated mathematics pathway that students begin, sometimes as early as sixth grade. Although students can exit before reaching the end at Calculus, there is no open entry along the way. Researchers have found that girls are more likely to drop out before completing Calculus (Lee & Ware, 1986; Moses et al., 1999; Oakes et al., 1990). Some of the reasons are poor grades, perceived lack of relevance, or lack of interest. For all students, the degree to which they like math is increasingly important to remaining in the pipeline as they move through (Burkam & Lee, 2003). School graduation requirements did not seem to play a role in remaining in the pipeline (Teitelbaum, 2003).

Ability grouping or tracking

This section reviews studies on the relationship between ability grouping or tracking and student achievement. Some researchers have examined the U.S. phenomenon from an international perspective. "The sorting of students in U.S. schools is so extensive and exclusionary that by grade 8 the proportion of students taking algebra is about the same as those taking advanced mathematics in grade twelve in other countries" (Useem, 1990, p. 1). In international comparisons, U.S. students do not seem well served by the current tracking practices. TIMSS results for grade 12 indicated American students were among the lowest of the 21 participating countries. NAEP (Main NAEP and trial assessment of the states in 1990) showed our best students, twelfth-graders intending to go to college and enrolled in the academic track, performed barely above the level required to successfully understand material introduced by seventh grade. Relatively few U.S. students seem prepared for advanced mathematics and U.S. students in general do not perform at an advanced level compared to students from other countries (Haury & Milbourne, 1999).

On the other hand, a recent study, sponsored by the College Board, administered questions from the advanced mathematics TIMSS 1995 exam to U.S. AP Calculus and AP Physics students who scored a three or better on their respective AP exam. The results showed that this representative sample of American AP Calculus students outperformed advanced or honors mathematics students in the US, and outperformed advanced students from each of the 18 countries that participated in the study (Gonzalez, O'Connor, & Miles, 2001). The complete report is available at http://apcentral.collegeboard.com/repository/ap01.pdf.ti_7958.pdf.

Sebring (1985) made the point that both aptitude and course taking contribute to performance. She concluded that the relationship between course taking and aptitude may be circular and not possible to separate and that having more of either or both (courses or aptitude) contributed to higher test scores. In an analysis of the relationship of verbal and mathematics abilities, she found there were very few students with high verbal/low mathematics abilities, indicating a strong positive correlation between verbal and mathematics performance. She struggled with the problem of disentangling ability and course taking.

> Without controlling for aptitude, one would overestimate the effects of coursework on test performance, since part of the effect would be the higher aptitude associated with students who take more coursework …The Educational Testing Service claims that aptitude scores capture both innate ability and school learning, so that aptitude scores can be viewed as both controlling variables and outcome measures. (p. 115, 120)

Researchers have come to different conclusions about the effect of tracking on students of different abilities. Kulik and Kulik (1984) found that acceleration had positive effects for high-ability students while other researchers contend that tracking hurts kids at the low-ability level and heterogeneous groups do no harm to students with high ability (Catsambis, Mulkey, & Crain, 2001; Oakes et al., 1990; Schoenfeld, 1994). Gamoran (1987) found that the difference in achievement between students in the upper and lower tracks was even greater than the difference between those who stayed in school and those who dropped out.

The year of Algebra completion

The primary method of enrichment in mathematics in American schools is to accelerate students and enroll them in a course in Algebra before high school. “The NCTM has emphasized the need for *all* students at the eighth grade to be taught a wide range of mathematical topics including estimation, functions, statistics, probability, measurement, and algebra" (Shakrani, 1996, ¶6). This NCTM recommendation does not necessarily indicate a formal course in Algebra, rather the teaching of algebra as one of many mathematical topics…"by removing the privilege status from early access to algebra, the argument suggests that schools provide a broader base for increased mathematical literacy for all students" (Smith, 1996, p. 142).

When a formal Algebra course is taught, there may not be a great deal of instructional time available to teach other topics. Researchers have surprisingly found that a course in algebra does not necessarily have a significant positive effect on performance on algebra items on some standardized assessments (Metcalf, 2002; Muthen et al., 1995). However, Ma (2000) examined six waves of data from the Longitudinal Study of American Youth (LSAY) and found that early high school Algebra significantly and positively affected achievement.

Smith (1996) found that early access to Algebra (before high school) “may ‘socialize’ a student into taking more mathematics" (p. 141). Although one of the reasons a student may take Algebra in grade 8 is to be on a path to take Calculus in grade 12, early access to Algebra does not guarantee that students will remain in a math course through all four years of high school. Studies have shown that over 30% of these

Algebra-8 students have stopped taking math after grade 10 (Partenheimer et al., 2001; Smith, 1996).

Smith cautions that policy changes to provide everyone with Algebra in grade 8 would probably dilute the effects. Algebra 8 might then be stratified to remedial Algebra, regular Algebra, expert Algebra or courses that "credential" students would then be Algebra seven, etc. She points out that the NCTM recommendation is that algebra *concepts* should be taught throughout grades 5-8. (NCTM, 1989, p. 102).

The NAEP results for Missouri and the nation are reported by course-taking levels. The data show that as higher percentages of Missouri students take grade 8 Algebra, the average score for that course-taking level decreases. The tables below indicate that increasing percentages of eighth-graders are taking a course in Algebra and fewer are taking General Mathematics.

Table 4

NAEP 8th grade results for 1992 mathematics assessment

	Algebra		**Pre-Algebra**		**Grade 8 Mathematics**	
	Percent	Average Score	Percent	Average Score	Percent	Average Score
Nation	19	299	28	271	50	253
Missouri	13	305	26	278	59	261

See: (http://nces.ed.gov/pubs96/web/96815.asp). (Shakrani, 1996)

Table 5

NAEP 8th grade results for 2000 (Main NAEP) mathematics assessment

	Algebra		**Pre-Algebra**		**Grade 8 Mathematics**	
	Percent	Average Score	Percent	Average Score	Percent	Average Score
Nation	25	301	31	270	37	264
Missouri	23	295	38	271	36	262

See (http://nces.ed.gov/nationsreportcard/mathematics/results/advanced-8.asp)

Students who enroll in Algebra in middle school are more likely to reach higher levels (37%) in the high school pipeline than students who do not take Algebra (29%) in grade 8 (Atanda, 2000). Also students with grade 8 Algebra and higher level courses in high school were more likely (72%) to enroll in a four year university than students who took the same high level high school course but did not take grade 8 Algebra (42%). Hall (2001) found that student SES was significantly and positively related to rigorous course-taking.

Horn and Bobbitt (2000) studied the influence of parents' level of education on student course-taking behavior in mathematics. The researchers speculated that parents' level of education influences the likelihood that parents will advocate rigorous course taking for their children. These researchers also learned that "when controlling on mathematics proficiency and parents' education, first-generation students (students whose parents did not complete college) increased their likelihood of completing advanced high school mathematics courses by taking Algebra in the eighth grade" (p. viii).

Educators have experimented with the concept of 'Algebra for all' with different results. Gamoran and Hannigan (2000) found that the benefit of taking high school Algebra is weaker for students with low test scores. The authors offered possible explanations for why low-scoring students might benefit less from Algebra taking. One is that they simply have less capacity to learn, another is that they are tracked into a less rigorous curriculum, and still another is that they are scheduled into regular Algebra classes where the instructional methods are not well suited to low achievers. Two

possible methods offered for providing access to Algebra for all were Equity 2000 (no longer an option) or what was referred to as a "stretch" curriculum that bridged the gap between general mathematics and Algebra by using an integrated hands-on approach (Gamoran & Hannigan, 2000).

The summary report for Equity 2000 (Harris, 1998) indicates that the participation rates in Algebra and Geometry increased at all of the sites. The passing rates for Algebra and Geometry may be interpreted as improving. Although the percent of students passing Algebra decreased at all the sites, since a greater number of students took the classes, the number passing increased in some cases. However, many students still failed despite increased efforts to provide support to struggling students.

The studies reported in this literature review used a variety of assessments as the dependent or criterion variable. There is no research in the literature that links MAP mathematics assessments to course taking, specifically the year of Algebra completion. There is also very little research (Metcalf, 2002) that links course-taking behavior to state-mandated, standards-based assessments like the MAP.

Graduation requirements

Researchers (Horn, 1990; Tuma & Gifford, 1990) studied course taking in mathematics and science for high school students from 1969-1987. The results of Horn's analysis showed that on average students in 1969 earned a high number of credits in mathematics and science. The average number of credits earned dropped from 1975 to 1982 and then increased from 1982-1987. This increase was observed for all types of students, regardless of gender or race/ethnicity. The increase coincides with the timing of

A Nation at Risk (NCEE, 1983), which urged more rigorous course taking. The NCEE recommendation was that *all* students should take the *Five New Basics* (National Commission on Excellence in Education, 1983). These new basics were said to form the core of the modern curriculum. This core included four years of English; three years each of mathematics, science, and social studies; and one-half year of computer science. The recommendation for college bound students also included two years of foreign language. The gender gap in course taking was closing in all but the highest level of mathematics courses where males still took more advanced mathematics classes. Although the study noted numbers of courses completed, it did not take into account the effect of these course-taking behaviors on student achievement. Horn recommended further research to evaluate the relationship between courses and achievement (Horn, 1990).

Schiller and Muller (2003) examined the relationship between high school course-taking behavior, state graduation requirements, and assessment and accountability policies. Although they discussed *No Child Left Behind* ("NCLB," 2002) and its accountability issues, their study used data from the early 1990s (NELS 88). They found that state graduation requirements had small but statistically significant effects on course taking, both on the types and number of courses. Even though students in states with higher graduation requirements tended to enter high school at a slightly higher level than students in other states, “students in states requiring more academic courses to earn a high school diploma tended to earn fewer advanced mathematics credits" (p. 9). The authors explain that this may be due to the fact that more courses are required in the other core areas, causing students to take more courses in those other subjects instead of more advanced mathematics courses (p. 29). **“**Students in states with a greater number of

academic courses required for high school graduation tended to be placed in higher courses as freshmen, to earn fewer advanced mathematics credits, and the influence of their freshman course placements was stronger in these states" (p. 10). Holding students accountable for test performance was correlated with a depressed number of advanced mathematics credits earned; whereas, increased accountability for test performance at the school level was the only strategy that seemed to increase all students' opportunities for learning mathematics in high school.

Teitelbaum (2003) used NELS 88 data to examine the relationship between graduation requirements in mathematics and science, course taking, and student achievement. He credited the development of graduation requirements in 41 states by 1984 to the call for more rigor in American schools sounded *by A Nation at Risk* (NCEE, 1983). The intent of the increased rigor was to have students study more mathematics and science to gain proficiency, especially those groups of students who were previously underrepresented in higher levels of mathematics and science (low SES, minorities). This study found evidence that schools requiring three or more mathematics courses had mitigated the influence of race on courses completed; however, the percentage of students with three or more years of high school mathematics completed still varied with track placement and grade 8 scores on the NELS: 88 assessment. A popular concern about increased graduation requirements was that there would be a greater number of lower-level courses offered to allow students to earn three credits without taking advanced classes. There was no evidence of dilution of courses in this sample. The most disappointing finding of this study was that the high school graduation requirement policies were not associated with student achievement. Student achievement in this study

was measured by the gain in scores from grade 8 to grade 12. However, the author stated that a limitation to the NELS: 88 data used in the study is that student achievement was measured by assessments that did not test any skills beyond the level of Algebra II. Therefore, the criterion variable in this case was not sensitive to higher-level mathematics course taking.

Tuma and Gifford (1990) found that all the growth in the average number of mathematics credits completed by non college-bound high school graduates in their study was at the basic or general levels, even during the period of reform. Only among the high school graduates planning to go on to a four-year college did the number of advanced mathematics and Calculus credits increase. This study raised some philosophical questions about whether the reform movement missed the mark. The increased graduation requirements did not seem to benefit the at-risk students they were aimed to help, nor did they have any affect on the college-bound students.

Alexander (2002) found that minority and poverty status of schools was related to course taking in urban settings and school size was a factor in all settings. After the reform in 1984, the percentage of minority students was negatively and significantly associated with the share of the curriculum devoted to advanced core courses and positively and significantly associated with greater shares of the curriculum devoted to noncore courses. The higher the rate of poverty, the lower the share of class time devoted to the core. Finn et al. (2002) also studied the effect of school characteristics (enrollment, urbanicity, and SES) as well as school policies (graduation requirements and course offerings) on course taking in the general student population. They found that increased graduation requirements seemed to benefit vocational students the most, general track

students somewhat, and had no effect on academic track students (p. 342). For all students, graduation requirements affected the number of mathematics courses taken but not the intensity (p. 364).

Arkansas graduation requirements call for three years of high school mathematics. Since many students complete their third course by the end of their junior year, they do not take a mathematics course as seniors (Berry, 2003). Researchers have recommended that all students take a mathematics course in each of the four years of high school (U.S. Department of Education, 1997). Berry went on to say that she believed that increasing the graduation requirements for mathematics to four years while allowing Algebra A and Algebra B (Algebra I over two years) to count for two years of credit defeats the purpose of the increase in required courses.

Clune and White (1992) studied the effect of changing graduation requirements (1982-1988) on course taking in all disciplines in four states, one of which was Missouri. They examined changes in course-taking behavior among graduates of high schools enrolling mostly lower achieving students in states adopting high graduation requirements in the 1980s. The criterion for inclusion in their sample was that a state sets requirements above the average of preexisting academic course taking. However, in the case of Missouri, the graduation requirements for mathematics did not change, nor have they changed since then. The graduation requirement for mathematics in Missouri remains at two years. Clune and White's data, compared to national data at that time, indicated that the course-taking trend was toward three plus credits of mathematics regardless of graduation requirements.

> The largest district gain occurred in a state with fairly typical mathematics requirement (Missouri, with a 2-credit mathematics requirement). This is one bit of evidence among many in our study that the state requirements are only one of the many influences on course taking in the high school curriculum. (Clune & White, 1992, p. 9)

Some of those other influences may be district requirements, state university entrance requirements, and possibly the entrance requirements of other universities. The most frequently added courses were those at the beginning of the college prep sequence rather than at the end. In mathematics, those courses were Pre-Algebra and Algebra I. The extra credits were a third of a year of extra mathematics and a half-year of extra mathematics in urban districts.

The freshmen admission requirements for the Missouri University system include four years of mathematics beginning with Algebra I. The university will accept an Algebra I course taken in grade 8, provided it is followed by Geometry in high school. The other Missouri State university campuses require three years of mathematics beginning with Algebra I but they strongly recommend four years.

Gender

Research related to the effect of gender on performance, especially in mathematics, will be examined. Studies related to gender issues will overlap with the issues of assessment and course taking. The studies reviewed in this section will be organized under these topics as they relate to gender: differential performance on

assessments, greater variability of scores for males, differential performance on item types, and differential performance on content strands.

Gender and assessment

Sex differences in mathematics achievement are well documented in educational research (Hyde et al., 1990; Maccoby & Jacklin, 1974; Wilder & Powell, 1989; Willingham & Cole, 1997). Researchers consistently find that males perform better than females on measures of mathematics achievement while females perform better than males on measures of reading and writing (Coley, 2001; Gambell & Hunter, 1999; Han & Hoover, 1994; Kleinfeld, 1998; Wilder & Powell, 1989; Willingham & Cole, 1997).

Willingham and Cole (1997) discussed the importance of disentangling constructs, cohorts, and samples (selective, representative, and available) when examining gender research. They define test fairness as comparability in assessment for all individuals and groups. The fairness issue is at the center of the gender and assessment debate. The intended use of assessments plays a role in the need to identify and intervene on behalf of both males and females where there are score differences by gender. Since important decisions are made about students and schools based on test scores, the study of differential gender performance is critical.

O'Neil et al. (2001) conducted a series of studies to determine if monetary incentives would influence student performance on low-stakes tests. In addition to finding that money did not motivate students on low-stakes tests, they reported other findings related to gender and assessment. An important finding in this study, and previous studies conducted by these researchers, was that in their particular sample (a

southern California population with a large percentage of LEP students) they consistently found that males outperform females on NAEP and TIMSS items. Although the aggregate data showed the mean male and female performance to be about equal, when they disaggregated the data by gender and items they found differential performance by gender and item. They made a strong case for studying the interaction effects of background variables and gender rather than using aggregate data to make the case that the gender gap is closing. Other researchers have also made the point that it is important to understand that group mean scores do not indicate the performance of all males or all females (Taylor et al., 1996).

Rebhorn and Miles (1999) wrote an essay in which they attempted to answer the question of whether the SAT-M is "the culprit or a magnifying glass" (p. 315) for gender differences. They, and others, see a problem with the SAT-M in particular because it is used as a high-stakes test for students. For seventh-grade students in the Talent Search program, the SAT is used to identify students with potential for success in postsecondary studies. The highest scoring students are given opportunities to attend special programs for high-ability students. The higher the cut score for participation, the greater the ratio of boys to girls in the qualifying group. The authors discussed the question of whether the differential performance on the SAT-M indicated a biased test or different opportunities to learn, without making any conclusions. The potential solutions they offered to the problem were focused on equitable use of the scores for participation in special programs rather than closing the gap in scores. They suggested that different cut scores could be used for boys and girls; the modified cutoffs would reflect the prevailing gender gap in scores for all test takers. They also recommended the use of multiple measures or criteria

for participation in special programs as well as more encouragement by educators and parents for females in mathematics.

A special subgroup of the Talent Search population, those who score between 700 and 800 on the SAT-M before age thirteen, began to be identified in 1980. This study of this subgroup of talented students, named the Study of Exceptional Talent (SET) was expanded to include students with exceptionally high verbal as well as mathematical talent. Between 1980 and 1992, 1,132 students, ages 8 to 13, qualified for SET. Females represented 18.9% of the SAT-M qualifiers, 55.5% of the SAT-V qualifiers, and 25.7% of the double qualifiers (Brody & Blackburn, 1996).

In 1991, the American Association of University Women (AAUW) produced a highly publicized report, *How Schools Shortchange Girls* (American Association of University Women, 1992). They made a strong case that girls are victims of a school system that causes them to fall behind boys in math and science. Kleinfeld (1998) claims the AAUW only told part of the story and that what the AAUW failed to include in their report was the evidence related to female superiority in reading and writing. Males lag behind females in reading and writing by far wider margins than the female lag in math and science. “The gender gap favoring females in reading and writing is more than twice the size of the gender gap favoring males in science and mathematics" (Gender differences in standardized tests of school achievement Section, Table 4). Kleinfeld was citing the *Digest of Education Statistics, 1997.* The most recent edition, 2002, shows the same differential performance by gender with girls far ahead of boys in writing, somewhat ahead in reading, and behind the boys' performance in mathematics. The margins favoring girls in reading and writing are significantly higher than those favoring

males in mathematics (National Center for Education Statistics, 2002b) (http://www.nces.ed.gov/programs/digest/d02/index.asp).

An examination of Missouri MAP data in Communication Arts shows that the gap that favors females persists and that the gap widens as students go through school. The number of females scoring Proficient or Advanced in Communication Arts in grade 11 in 2003 was over twice the number of males (http://dese.mo.gov/divimprove/assess/spring03modisaggregatemaptotals.html).

The differential performance of males and females continues to be a controversial topic. Researchers have posited biological reasons for superior male mathematics performance (Benbow & Stanley, 1980; Halpern, 2000). Females' interests and attitudes toward the study of mathematics have been offered as explanations for course-taking behavior of females as well as for the lower scores girls earn in mathematics assessments (Ayalon, 2002; Oakes et al., 1990; Thorndike-Christ, 1991; Willingham & Cole, 1997). This may be an effect of acculturation because a cross-cultural study done by Feingold (1994) found that girls outside the United States do not dislike math and science the way American girls do. In the same study, he found that in some countries girls scored better on spatial tasks and in other countries boys did better, indicating that a purely biological explanation is not likely to account for all of the gender difference in math performance particularly on spatial tasks. Cooper and Dunne (2000) studied mathematics performance of students in the United Kingdom at ages nine and 13. They were particularly interested in the interaction effects of ability, gender, and social class. Their results showed that gender did not appear to play a statistically significant role at the secondary level. A study set in Canada found that girls have outperformed boys at the end of grade 12 in

Saskatchewan in mathematics and science as well as reading and writing since 1987 (Gambell & Hunter, 1999).

Variability

A major and consistent finding in the study of gender differences in quantitative ability is that male test score distributions have greater variance than female test score distributions (Beller & Gafni, 1996; Bevan, 2001; Fan & Chen, 1997; Feingold, 1992; Halpern, 2000; Han & Hoover, 1994; Hedges & Friedman, 1993a, 1993b; Hedges & Nowell, 1995; Kleinfeld, 1998; Maccoby & Jacklin, 1974; McKendree, 2002; Nowell & Hedges, 1998; Willingham & Cole, 1997).

Gender differences are often reported in terms of effect size, d; where d= (Mean of the male scores-Mean of the female scores)/pooled within group standard deviation (Cohen, 1969). Cohen defined categories for significance of these effect sizes: .20 to .49 is a small difference, .50 to .79 is a medium difference and .80 and above is considered a large difference. Some researchers reverse the order of the male and female means in the numerator but all effect sizes reported in this literature review will be reported with positive values of d favoring males. The relative variance of male and female distributions is reported in the research as Var_m/Var_f, so values greater than one indicate greater variance for males.

Feingold's 1992 study examined gender differences in variability on several standardized test batteries. He concluded that males were consistently more variable in quantitative reasoning, spatial visualization, spelling, and general knowledge. However, these differences in variability were coupled with differences in means. He recommended that researchers look at both differences in variability and central tendency in order to

make conclusions about gender differences in cognitive ability. Feingold demonstrated that both effect sizes (d) and variance ratios (VR) are moderated by the year the test was normed, the grade of the examinees, and the interactions between those two factors. Hedges and Friedman (1993b) reexamined the results from Feingold's study and disagreed somewhat with his statistical processes but applauded the valuable contribution Feingold made in emphasizing the need for studying both effect sizes and variance ratios in the study of group differences.

Hyde et al. (1990) conducted a meta-analysis of over 100 studies looking at a total of over 3,000,000 subjects from age five to adulthood and yielding 254 independent effect sizes. Because SAT subjects represented over 20 percent of the original sample, they had a disproportionate effect on the mean effect size. The researchers excluded the SAT results from the total results for this reason and analyzed SAT studies separately. They used General Linear Modeling (GLM) to determine significant predictors for factors contributing to effect size. The three most significant predictors, in order of their magnitude, were age of the subjects, selectivity of the sample, and the cognitive level of the test. In general samples, they found a non-significant overall effect size of –0.05 favoring females. General samples had the characteristics of mixed or unreported ethnicity, mixed or unreported cognitive level, and mixed or unreported mathematics content. In all samples (except SAT), they found a 0.15 effect size, favoring males. They found a moderately significant effect size of 0.29, favoring males, for problem solving in high school aged students.

Another study that examined norm-referenced test scores supported the gender variability findings. Han and Hoover (1994) studied results of administrations of: Iowa

Test of Basic Skills (ITBS), Iowa Test of Educational Development (ITED), and Tests of Achievement and Proficiency (TAP) from 1963-1992. They found "the nature and magnitude of the differences between male and female scores have remained similar over the last 30 years. Females have consistently scored higher than males in Reading Comprehension and Language Total" (p. 6). On the topic of variability, the differences in means were small but there was an advantage for above-average males and below-average females. As Han and Hoover analyzed scores at the tenth, 50^{th}, and 90^{th} percentiles, they suggested that it was not surprising that many researchers noted gender differences in achievement. Many studies use data from highly-selective samples (SAT, college-bound) and evidence shows greater variability for males so there are more males at the top and bottom. In examining selective samples, the sample is skewed in favor of males at the top end. They made a recommendation that a critical policy for educators is to plan interventions to assist males who perform poorly in language and reading.

Studies demonstrate a link between item difficulty and gender. This link is very complex. Because of other research findings of greater variability in male scores, Bielinski and Davison (1998) hypothesized that females would perform better on easier items and males would perform better on more difficult items. Their data supported this hypothesis. They followed with another study using data from eight different populations. The data included student performance on multiple-choice mathematics questions from 1992 NAEP, TIMSS, and NELS: 88. They studied the relationship between (item difficulty $_{male}$- item difficulty $_{female}$) and (item difficulty $_{total}$) and found the same negative correlation, indicating "easy items are easier for females than males and hard items are harder for females than males" (Bielinski & Davison, 2001, p. 51). These researchers

cautioned that other studies (e.g. Harris & Carlton, 1993; Lane et al., 1996; Ryan & Fan, 1996) that point to an observed gender by item interaction or gender by content interaction should also examine item difficulty as a factor that can explain differential performance of males and females. There is an additional confound between item and examinee characteristics. On several norm-referenced tests, the average item difficulty decreases as the grade level on which the test was normed increases. The process for determining item difficulty on these tests was the proportion passing the item. (Bielinski & Davison, 2001, p. 52) This has important implications in interpreting differential performance over time to indicate a growing gender gap as students get older. "The size and direction of the achievement gap…may arise from differences among students…items…or both" (Bielinski & Davison, 2001, p. 53).

Nowell and Hedges (1998) examined trends in gender differences from 1960-1994 by analyzing means, variances, and extreme scores from eight national samples of twelfth-graders. In analyzing the data they examined the proportion ratio in the tails (relative proportion of males/relative proportion of females). They also partitioned the gender differences into the portion resulting from difference in means and the portion resulting from difference in variance. In general, they found larger differences in variance were correlated with larger mean differences (correlation between all computed variance ratios and means was 0.74). They found that the "gender differences in mean and variance are small, while differences in extreme scores are often substantial" (p.2). They predict that the difference in means for males and females will disappear in 40 years but the proportion ratio by gender in the tails when the means are congruent will not be

equal. They predict that in the 90th percentile, there will be a 1.2:1 ratio of males to females and in the 95th percentile, there will be a 1.3:1 ratio of males to females.

In a study using a large sample (23,000) and three waves of data from NELS: 88, males had only a slight advantage over females in the total population (Fan and Chen, 1997). However, at the high end, males outnumbered females by a considerable margin. The higher the achievement level they examined, the smaller the percentage of female students they found. Male students outnumbered females by two to one in the 95th percentile group in twelfth-grade. This study also found greater variability in the distribution of math achievement scores for males. Rock and Pollack (1995a) explained the structure and methods of NELS: 88. Unlike NAEP, which is cross sectional, NELS: 88 is a longitudinal data set that is designed to measure growth in achievement of a cohort over time. Studies such as NELS: 88 "are important from a policy viewpoint because they provide information on the relationship between gains in achievement and course-taking behaviors" (p. 1). The 1988 eighth-graders were tested three times at two-year intervals; once in grade 8, again in both grades 10 and 12. The content in the NELS: 88 mathematics tests spanned topics from basic mathematics through Algebra II, but did not go as far as Precalculus. "The seven test forms were put on the same scale so that comparisons of scores between time points could be made" (p. 1).

Item type

Researchers emphasize that it is important to understand that items that discriminate are not necessarily bad items. (McKendree, 2002; Willingham & Cole, 1997; Rebhorn & Miles, 1999). Ryan and Fan (1996) stated that when items have Differential Item Functioning (DIF), the recommendation is not necessarily to remove

them. If the content integrity of the test is at issue, the items should not be removed. Instead, DIF items may have implications for curriculum, rather than assessment.

Wilson and Zhang (1999) studied differential gender performance on constructed response and multiple choice mathematics items on a 1995 Delaware state assessment, given at grades 3, 5, 8, and 10. The state gave two assessments to students. One was an "interim" assessment that consisted of 10-15 CR items. The items all focused on a common theme and a single content strand (such as number sense or geometry). The second assessment was a norm-referenced MC test used primarily for Title I reporting. The researchers sorted the test items from both assessments, at all grade levels, into one of three categories defined by the "mathematical processes described in the NAEP framework: procedural skills, conceptual understanding, and problem solving" (p. 5). They were especially interested in students' performance on items that required the students to communicate their mathematical thinking. They sorted the items a second time, based on whether or not the item demanded communication of mathematical ideas and found that none of the MC items required such communication. The rationale for the second sorting was that "language arts skills have often been seen as a particularly female strength, so items that assess these skills would be of particular interest in a study of gender differences" (pp. 6-7). They concluded that while the gender gap may be narrowing on MC items, it is still present on items requiring communication. Their results showed that males outperformed females on CR items and in the area of problem solving. The gender gap in communication and problem solving increased with the age of the subjects.

Hamilton (1999) conducted a study analyzing results for twelfth- grade science CR items using NELS: 88 data. Using Logistic Discriminant Function Analysis (LDFA), she found one CR that displayed a large male advantage, contributing to the gender difference on total score. The item involved spatial mechanical (SM) reasoning. Male students scored an average of nearly one half standard deviation higher than female students in the SM dimension. Hamilton supplemented the quantitative data with interviews of 25 high school students. Using both quantitative and qualitative data, Hamilton made the point that the observed difference in performance, favoring males, on SM items may be due to knowledge and skills acquired outside of school.

Pomplun and Sundbye (1999) studied gender differences on CR reading items for 500 seventh and tenth-grade students on a large-scale state assessment in Kansas. They considered several factors to account for differential performance of males and females including difference in reading skills and length of responses (number of words written). They found that the gender difference in grade 10 was statistically significant (d = -0.39) in favor of females and that reading ability did not totally explain the differential performance. In 10^{th} grade, the correct answer accounted for most of the score difference followed by the number of words written. The length of the response accounted for the largest decrease in gender score differences. The number of words in the response was categorized as both construct relevant (providing additional information) and irrelevant (demonstrated greater effort). These and other researchers have expressed concern about the validity of CR and PE items because raters can be influenced by construct irrelevant factors. In Missouri, the student responses are not returned so it is not possible to determine whether or not students are including irrelevant information in their responses,

or more importantly, whether this information is affecting the score they are given for their response.

Lane et al. (1996) examined gender-related differential item functioning (DIF) on 42 middle school mathematics performance assessments that were each administered to approximately 500 sixth-grade or seventh-grade students. They used logistic discriminant function analysis (LDFA) and found two tasks that favored males and four tasks that favored females. Although previous research has found that tasks embedded in a real-world context favor males, that was not the case in this study. One of the items that favored males and all of the items that favored females were in a real-world context; however, the majority of the 42 tasks were set in real-world contexts. The authors reported that all students in the sample received instruction in standards-based classrooms where real-world applications were emphasized so the context factor may not have been a critical one in this sample. They found a significant gender difference with respect to showing work to support your answer. On a ratio and proportion task that favored males, they found that males were less likely (83% of males showed their work and 93% of females) to show their solution strategy; however, when males did show their work, it demonstrated the use of an appropriate strategy more often than females (47% males to 34% females) and males more often arrived at the correct answer (30% males to 19% females). Girls were also much more likely to provide complete work (82% females vs. 55% males). In a number sense task that favored females, the girls tended to provide more conceptual explanations than boys did. Since Missouri educators do not have the opportunity to examine the scored work of students on the MAP, it is not possible for

them to use specific student responses on the MAP to guide instruction on open-ended performance assessment tasks.

McKendree (2002) stated that current evidence does suggest a weak systematic gender bias where MC questions favor males and CR questions favor females. Anderson (2002) conducted a study on open versus closed assessment tasks using three groups of college-level students. He found evidence of differential gender performance by item type, finding that women performed poorly on open-ended items. Because the women in the sample were said to be the top three or four percent of their cohort, he concluded that this differential performance was related to affective factors rather than knowledge. There was some evidence that girls were reluctant to take risks when they were uncertain and there were stated accountability factors related to performance. However, the study's statistical methods may have been flawed. There was a 3 to 1, male to female ratio in one group, and a 2.5 to 1 male to female ratio, in each of the other two groups. In addition, not all subjects were given identical tasks.

The SAT and PSAT were revised in 1993-1994 in order to (a) better align with NCTM standards, (b) align with current cognitive theories of learning, (c) provide more useful feedback, and (d) respond to threats to score validity (coaching, guessing, speed of response) (Burton, 1996). The changes were also prompted by concerns that teachers were trying to teach to a test that was traditionally not curriculum-based. In addition, researchers have noted that females' test behaviors indicate low tolerance for risk-taking and slower response times. The actual changes in both tests, with respect to the NCTM standards, involved setting more problems in real-world contexts; adding more problems that require interpretation of statistics, graphs, and tables; and allowing the use of

calculators. Computation was de-emphasized in the revised test and CR items were added. In addition, the time allowed for the mathematics test was increased by 15 minutes. Males have traditionally scored higher than females on the SAT-M and females have outperformed males on the SAT-V. The revised test showed a gain in verbal scores for females and no change in mathematics, but the researcher cautioned that it would require several years of data to determine the effects of the changes.

Myerberg (1996) used 6 different assessments in one school district to examine relationships between assessment types (MC, short answer, extended answer), gender, and racial and ethnic group membership. He found that non-multiple choice tests in mathematics, language arts, and reading favored females and the trend was reversed for males. Myerberg cautioned that the study might not be generalizable because it was conducted in only one school district and only used these six assessments, some of which were locally-developed.

Content strands

Hyde et al. (1990) were frustrated in their attempt to identify the mathematical content of the tests in their meta-analysis. “We must know if there are large gender gaps for certain types of content. That can be determined only when researchers construct tests and report results that assess the various kinds of mathematics content separately" (p. 155). They recommended looking at other factors to explain fewer women in college mathematics and mathematics careers: i.e. pre-college curriculum, attitude, and sex discrimination in education and employment.

Armstrong (1981) stated that differential achievement is not solely a function of course taking. Females performed slightly better than males at age 13 in algebra and spatial visualization tasks. However, by twelfth-grade, even with course taking controlled, girls had lost their edge in those two categories and the differential performance by gender in problem solving had increased. Males consistently outperformed females in problem solving and the gap between them grew as their age increased. A meta-analysis of gender differences in spatial ability (Linn & Peterson, 1985) found that the magnitude of the gender difference depended on the type of spatial ability being tested. Spatial perception ($d = 0.44$) tasks, spatial visualization ($d = 0.13$), and mental rotation ($d = 0.73$) all had differences that favored males. "In spatial perception tests, subjects are required to determine spatial relationships with respect to the orientation of their own bodies, in spite of distracting information" (p. 1482). An example would be drawing a horizontal water line in a tilted bottle. Mental rotation assesses the subject's ability to rotate a two or three-dimensional figure rapidly and accurately. Items "are used to measure the time required rather than the accuracy of solution (which is extremely high)" (p. 1484). Spatial visualization tasks "involve complicated, multistep manipulations of spatially-presented information. The tasks 'may involve the processes required for spatial perception and mental rotations but are distinguished by the possibility of multiple solution strategies'" (p. 1484). Examples are embedded figures, where subjects must find a simple shape in a complex shape; or paper-folding tasks, where subjects must choose how a folded paper would look when it is unfolded.

Brosnan's (1998) study of male/female performance on spatial tasks revealed that attitudes affected performance. Brosnan found that females' performance was poorer when the task was described as a measure of spatial ability. Females outperformed males within these areas (traditionally male areas) when the questions were part of a compulsory aspect of education. This finding undermines the notion that a stable sex difference in spatial ability represents underlying causality (Brosnan, p. 205-206).

Friedman (1995) examined the relationship between spatial and mathematical skills by conducting a meta-analyis of correlations of spatial and mathematical tasks from studies of K-12 and post-secondary subjects. In order to put these correlations in context, Friedman also considered correlations between verbal and mathematical measures. The meta-analysis showed that "when space-mathematics correlations are combined and compared to other correlations, they are not convincing evidence that spatial skill is well-related to mathematical ability" (p. 40). The findings indicated higher correlation between mathematical and verbal ability (0.35 to 0.57) than mathematical and spatial ability (0.30 to 0.45). There is evidence that as the selectivity of the sample becomes greater (gifted or college-bound), the mathematical-spatial correlations for females are higher than males.

Taylor et al. (1996) studied students participating in a mathematics competition in Australia from 1983-1992. The test consisted of only multiple-choice items. They found that the overall gender gap in mathematics is closing, but their findings showed the gap is increasing, favoring males, in questions categorized as algebra. This is contrary to the reports of other researchers who found that that girls performed better than boys on algebra tasks (Bevan, 2001; Harris & Carlton, 1993).

In the Lane et al. (1996) study, the only performance assessment item that showed severe DIF favored males and it was a geometry item. Students were required to provide reflections for a given figure. In fact the two performance items that favored males both contained a figure and required no verbal explanation. None of the four items that favored females contained a figure. Two of the four items favoring females required a written verbal response. Willingham and Cole (1997) also found that men performed better on items containing a figure.

In a review of research on gender differences in mathematics, Bevan (2001) found evidence that boys performed better in measures, rate, and ratio. Bevan also found there were more boys at higher achievement levels and that girls performed better in whole numbers, decimals, and slightly better in algebra.

An international study of gender differences for nine and 13 year olds in mathematics and science revealed differences overall on the subdomains (Beller & Gafni, 1996). Boys performed better on measurement and problem solving. Their study joined others that found the score variance was greater for boys than girls. Preece et al. (1999) in a UK study of 14-year-old students' science achievement found that males performed better on questions that discriminated well and on questions requiring interpretation of two-dimensional diagrams of three-dimensional phenomena.

Harris and Carlton (1993) examined gender differences in performance on the SAT-M by matching students by overall score and then looking for patterns in DIF both by item type and content. They also examined the subject matter in which an item was embedded. Over half of the categories of items considered yielded no statistically significant differences in gender performance. However, when overall scores were

matched the following differences were significant. Males performed better on geometry, items with visual stimuli, items that were related to real-life applications, non-textbook type items, items involving special topics such as money, time, rate, measurement, percents, averages, and areas; people items, longer items and items involving more difficult reading. Females performed better on algebra items, items that involved lower-level reasoning, general solutions, fractions, counting, unapplied mathematics, items with variables, and textbook-like items.

Interaction effects

Course taking and gender

McLure, Boatwright, McClanahan, and McLure (1998) examined the relationship between trends in ACT Mathematics scores and mathematics course taking from 1987 to 1997. The results showed that course taking accounted for 34.5% of the variance in scores. When examined by gender, course taking accounted for 33.9% for females and 36.7% for males. This is interesting since, in this sample, girls surpassed boys in 1990 in the average number of years of mathematics. The average mathematics credits for girls went from 2.97 in 1987 to 3.53 in 1997 while boys went from 3.07 to 3.44. This sample is somewhat selective because these are all students who took the ACT, who were likely to be college-bound students. Although the girls took more mathematics courses, those courses accounted for less of the variance in ACT mathematics scores. It may be that while the average number of courses taken by girls was greater than the number taken by boys, the level of difficulty of those courses may have been different with boys taking

more of the advanced courses. The data from their study is not clear on the intensity of the courses taken; just the number of credits earned.

Using the 1990 National Assessment of Educational Progress (NAEP) transcript study to analyze course-taking behavior, researchers found that males predominate in most-advanced and least-advanced courses while overall the genders earn about the same number of Carnegie units (CU) of mathematics credit. On average, students earned 3.11 CUs, a little more than the three units recommended by the National Commission on Excellence in Education (NCEE) in 1983 (Davenport et al., 1998).

Pallas and Alexander (1983) examined the differential coursework hypothesis by examining SAT-M scores for 6,119 twelfth-graders who were subjects in a longitudinal study, *ETS's* (Educational Testing Service) *Study of Academic Prediction and Growth*. They constructed regressions of SAT-M on combinations of background variables, coursework, grade-point average (GPA), and ninth-grade scores on the School and College Ability Test-Quantitative (SCAT-Q). ETS describes the SCAT as a measure of school-learned ability, designed to gauge a student's preparation for the next highest level of schooling (Alexander & Pallas, 1984, p. 399). They found the ninth-grade scores accounted for more than four times the variance than the next highest independent variable (level of education of the father). The ETS longitudinal study began collecting achievement data when the students were in grade five. At that point, the mean scores of males and females were equal. The reported mean difference for males and females on the SAT-M in 1968 (twelfth-grade) was 36.78 points, in favor of males. When the authors examined gender differences with a regression that included parents' level of education, race, sex, and grade nine SCAT-Q, the mean score on the SAT-M for males

was 35 points higher than females. When they added in coursework, the difference was 14 points in favor of males. When they added in GPA, the difference between mean scores for males and females on the SAT-M increased to 20.5 points. So controlling for GPA increased the residual female shortfall by 6 points. Females earn higher grades, so if the female grades were not better than male grades, the SAT-M M/F gap would be larger.

Hedges and Nowell (1995) attempted to use large data sets to examine differences in ability by gender. Large differences in verbal ability favored females. These researchers mention that gender differences are frequently attributed to differential curriculum. They also state that this differential curriculum is unlikely to be the cause in their study since all students are taught writing skills. They found that girls consistently performed significantly better at writing over the 32-year period covered by the data in their study

Girls have better grades than boys but lower test scores. There are at least two possible explanations for this. One is that the tests are biased against females (Miller & Mtchell, 1994). The other is that teachers' grading criteria and expectations are more consistent with girls' behavior; therefore, girls earn higher grades (Kleinfeld, 1998). One or both of these factors in combination may be the cause for the discrepancy. Wentzel (1988) found that girls' GPAs are related to social competencies that depend on cooperation with adult authority. Young (1994) found that females' college grades are underpredicted (related to course selection for females), and minorities are overpredicted. Possible explanations offered for higher female grades were that courses and departments with higher average grades have a higher proportion of women enrolled.

Wainer and Steinberg (1992) did a bi-directional validity study using SAT-M as the dependent variable with courses and grades as the independent variables in one case (retrospective study). In the other case (prospective study), they used SAT-M, gender, and gender by SAT-M as the independent variables with grades in the first year college mathematics course as the dependent variable. The retrospective study showed that males who take Calculus first semester have a 38-point advantage over females on the SAT-M, whereas the prospective says the males have a 64 point advantage. This is an example of scores on the SAT-M underpredicting girls' grades.

Course taking, gender, and assessment

One recent study (Metcalf, 2002) examined the relationships among the intended, implemented, and attained curriculum, using a sample of 3,019 10th grade students from three Illinois school districts. Four course-taking groups (based on the highest level of course enrolled in by the time of the 10th grade test) defined the intended curriculum or tracks: General Mathematics, Algebra, Geometry, and Advanced. The implemented curriculum was determined from a group interview with three high school mathematics department chairs from the largest of the three participating school districts. The attained curriculum was determined by student achievement, as measured by the 1998 10th grade Illinois Goal Assessment Program (IGAP) mathematics test. The Illinois Department of Education provided individual student's responses for this sample to the researcher. A major focus of the study was to examine differential item functioning (DIF) as well as differential bundle functioning on the seven IGAP goals (DBF) for the four curricular levels. The results showed that "the more mathematics the students took, the higher they

scored" (p. 47). He also found that when controlling for race and course-taking, higher SES students had higher scores.

However, an interesting finding was that the high-performing low track students outscored the low-performing high track students, which Metcalf interpreted as strong evidence that these low-tracked students could succeed in the higher track classes. The data suggested that track placement was related to SES and race but not gender. All groups had the weakest performance on the IGAP goal on measurement. An analysis of the percentage of item content covered by each group showed that the General group only received instruction in 54.3% of the total item content, and 0% of the measurement item content. The Algebra group received instruction in 60% of the total content and only 10% of the measurement content. Both Geometry and Advanced groups received instruction in over 95% of the total item content. An unexpected finding was that the Algebra group only covered 60% of the algebra items.

In some cases, the students in the lower groups performed better on measurement items than the students in the upper groups. The department chairs hypothesized that these students from the General Math and Algebra tracks may have learned the skills required for those items somewhere other than in their mathematics classes. Although the state department did not find DIF when they examined test items for racial or gender bias, Metcalf did find DIF on 47 of 70 items, when he examined test items based on curricular groups. When there was discordant coverage of item content in the curriculum "the odds were greater than 3:1 that the item would favor the reference group" (p. 97). One conclusion Metcalf makes is that OTL-DIF items and OTL-DBF item bundles address potential weaknesses in the curriculum offering. He recommends further research to

"consider whether test builders should create tests that are designed with OTL-sensitive DIF items and DBF item bundles, if it results in improvement of the mathematics curriculum and opportunity-to-learn" (p. 106).

Summary of the literature review

Both the use and the interpretation of student scores on various assessments determine what the relationship should be between the tests and the curriculum. Tests that are used to determine student access to college, to scholarships, or other special programs have high stakes for students. Therefore, it is critical for students' success that they are offered the opportunity to learn the mathematics that will allow them to succeed on these high stakes tests. There is ample evidence that girls get better grades and boys get better scores on such tests as the SAT-M. As long as multiple measures are used to predict student success in programs or colleges, the differential performance may not be a problem. Because of greater interest, ability, or different test-taking behaviors, males may perform better in problem solving or spatial reasoning and girls may perform better in verbal tasks. Differential item functioning may have implications for the curriculum, rather than the tests. The real problem lies in differential performance because of denied opportunities to acquire needed skills.

The research consistently shows that taking more mathematics courses and more rigorous courses leads to greater levels of post-secondary success, as well as higher test scores. Although the current trend in assessment is to move away from strictly norm-referenced multiple-choice tests, most of the research that looks at correlates of student achievement uses data from multiple-choice tests. The research on constructed response

and performance items has not effectively shown that these types of questions are valid measures of student achievement. Efforts to use external measures to validate performance tests are nil. The use of test data becomes more complex when the items are open-ended. When the actual item and response is not available it is difficult for educators to use the data prescriptively to modify curriculum or instruction. Group means have been shown to mask important differences of special groups within the population. These "sub-groups" have become more important than ever now that the national accountability model of NCLB demands proficiency for the aggregate and the disaggregated groups.

The effect of *A Nation at Risk* seems to be that it raised graduation requirements but did not change the course-taking behavior of the college-bound students. Although more credits have been required and earned in mathematics, the additional credits have been at the lowest levels and have not translated to increased student achievement. So the reform movements have not helped the group that was the target of the reform, those at the low end of course taking and achievement.

The greater variability of male scores is difficult to explain. It could be a result of greater interest and ability at the highest levels for males. In addition, males have been found to have more efficient test-taking strategies on norm-referenced multiple-choice tests where time is a factor. The variability data comes from such tests. The greater number of boys at the low end may be consistent with the greater numbers of boys who have historically lost interest or motivation and dropped out of school.

The distribution of MAP scores is not known. Since it is a criterion-referenced test, it is not desirable for the scores to be normally distributed. The state has expressed a

goal of having increasing percentages of students at the top two levels. Since Missouri's NCLB plan states a goal of having 100% of our students scoring Proficient or above by 2014, the percentage of students at the top two levels must increase each year. Although MAP results can be analyzed by content strand and item type, it may not be possible to analyze differential performance by gender on something as specific as spatial tasks. Without the specific item to examine, the item may only be able to be placed in a broad category, such as geometry.

Another factor related to testing is time; teachers put time into preparing students for tests and schools sacrifice many hours of instructional time to the administration of these tests. The data are returned when teachers have moved on to a new group of students and students have moved on to new classrooms. Most teachers do not have time to devote to a detailed analysis of test data belonging to students they no longer teach. Many school districts do not have sufficient funds to allocate staff to the task of analyzing the data. If someone in a central office analyzes the data, they do not have the benefit of understanding the students and the instructional practices of their teachers. The MAP test is a low-stakes test for individual students and high stakes for schools, districts, and states. It is important to learn what we can about which educational practices can influence student achievement on this test.

This chapter included the review of literature related to mathematics assessment, course-taking behavior, gender studies, and interactions between these factors. Chapter III outlines methods that were used in collecting and analyzing the data for this study, information about subjects, research design of the study, instruments, procedures, and human subjects concerns. The results of the statistical analyses will be reported in

Chapter IV. Chapter V includes a summary of the study, discussion of the research findings, conclusions, implications of the findings, and recommendations for future research.

CHAPTER III

Introduction

The purpose of this study was to examine the relationship between course taking in mathematics and achievement scores on both the Missouri Assessment Program (MAP) in grades 8 and 10, and the TerraNova in grades 8 and 9. This chapter contains a discussion of the methods used to collect and analyze the data in this study. This chapter contains sections on subjects, research design of the study, instruments, procedures, data analysis, and human subjects concerns.

Subjects

The focus of this study was one suburban school district in east central Missouri because of the following factors: the availability of the data to the researcher, the demographics of the district which are near the average of the state and county, the continuous enrollment of students for three years in the same district with the same exposure to the grade 8 through grade 10 mathematics curriculum. The demographic information for the district, county, and state is taken from 2003 data reported by the Missouri Department of Elementary and Secondary Education (DESE). (www.dese.mo.gov/schooldata/). The quartiles are reported where the first is the lowest and the fourth is the highest. The per-pupil expenditure for the district was approximately $7,000. This places the district in the first quartile for the county and the third quartile for the state. The median per pupil expenditure was $8,455 for the county and $6,536 for the state. The racial composition of the district is predominantly White, with minority representation of approximately 17%. This places the district in the first quartile for the county and the fourth quartile for the state. The median percent of minority students was

32.4% for the county and 2.3% for the state. The percent of students eligible for free or reduced lunch was approximately 24%. This places the district in the second quartile for the county and the first quartile for the state. The median percent of students receiving free or reduced lunch was 31.98% in the county and 45.8% in the state. The percent of residents in the participating school district with less than a high school diploma or equivalent was approximately 9.0%. The median household income in the population served by the participating school district was approximately $40,000 in the year 2000.

The number of students attending kindergarten through grade twelve in this district was less than 5000. This district has one middle school and one high school. All of the teachers who were teaching mathematics in the middle school and high school during the years of the data collection were fully certified and would be considered "highly qualified" by NCLB standards ("NCLB," 2002). The MAP mathematics scores (percent of students at the top two levels) in grade eight were in the second quartile for the county for grade eight in 2003 and the third quartile for the county for grade 10 in 2003.

The participants in this study were students in a small, suburban Missouri school district. The test scores of all students who

1. were enrolled in grade ten in this district in 1999-2000, 2000-2001, 2001-2002, 2002-2003
2. took the MAP mathematics test in grades 8 and 10, and
3. took the TerraNova test in grades 8 and 9 were used in the study. The sample was drawn from six years of test data (between 1998 and 2003) using four cohorts of students

Student Cohort	Grade 8 MAP	Grade 9 TerraNova	Grade 10 MAP	n
A	1998	1999	2000	207
B	1999	2000	2001	185
C	2000	2001	2002	223
D	2001	2002	2003	188

Students with valid scores in each data field were retained in the sample. This reduced sample size. A subset of the total sample was used for the ANCOVA in grade eight because TerraNova Communication Arts scores were not available for Student Cohort A. Only Student Cohorts B, C, and D were used in that one analysis. All other samples contained all students in all four cohorts with valid scores in each data field.

The data are archival and were accessed through a combination of the following resources: Student Information Systems (SIS), an electronic database that includes student transcript information; permanent record files of students; district reports that contain individual scores on TerraNova and MAP; Clear Access, an electronic reporting system of MAP scores; and TestMate Clarity, an electronic reporting system of TerraNova Scores. Permission for the use of the data was secured in writing from the school system superintendent. After test scores were matched to student transcript and demographic information, the data for each student record was entered into a spreadsheet that was used to import data into Statistical Package for the Social Sciences (SPSS), Version 13. Each student name was matched with an identification number assigned by the researcher. Only the researcher knows the names of the students that correspond to the identification numbers. Each individual student's data was linked only to the

Identification number in the spreadsheet file and in SPSS. Neither individual students nor individual schools were identified in the study, so no further level of permission was required.

Instruments

The data collection instrument in the appendix was used. The primary researcher completed the data collection form for each individual student. All data came from archival records of the school district. Gender, ethnicity, and transcript data were gathered from the Student Information System (SIS) and the Missouri Assessment Program (MAP) scores from the student record in the permanent files or from Clear Access for the years when that data-reporting system was available. TerraNova scores were collected from the student's permanent record, from district reports of student scores, or from TestMate Clarity. TestMate Clarity is an interactive CD-Rom report of student scores on TerraNova, one of the forms of score reports purchased by the school district from CTB McGraw-Hill.

Student achievement was measured by grade point average (GPA) in mathematics courses for grades 8, 9, and 10 (using the average of the semester grades); MAP–8 and MAP-10 mathematics scores; TerraNova Level 19-grade nine Mathematics scores; and TerraNova Communication Arts scores for grades eight and nine.

In Missouri, the Outstanding Schools Act of 1993 required the development of state standards and an assessment to measure student mastery of those standards. The assessment developed is the Missouri Assessment Program (MAP). The standards are the Show-Me Standards. There are 73 standards, 33 performance (process) standards and 40 knowledge (content) standards. The six content strands in mathematics are: (a) Number

Sense, (b) Geometry and Spatial Sense, (c) Data Analysis and Probability, (d) Patterns and Relationships, (e) Mathematical Systems, and (f) Discrete Mathematics.

The Missouri Assessment Program (MAP) was developed as a criterion-referenced, performance-based assessment system used to measure student progress toward mastery of the Show-Me Standards. The first year that all Missouri public school students were required to take the MAP mathematics tests in grades 4, 8, and 10 was in 1998. The MAP contains three sessions and three types of items. The entire test (all three sessions) takes approximately 3 to 5 hours to complete. Only the Multiple Choice session is timed. The Missouri Department of Education contracted with CTB McGraw-Hill to support the development and administration of the MAP. The Multiple Choice (MC) component is the Survey portion of the TerraNova, a nationally norm-referenced achievement test published by CTB McGraw-Hill. The Constructed Response (CR) items require students to supply an answer and in some cases to show their work and explain. The Performance Event (PE) items not only measure students' knowledge but also their ability to apply that knowledge to complex real-life situations. Students can be expected to work through a multi-step process, justify their solution, and provide explanations that include showing and labeling their work. These are more complex problems and there can be multiple acceptable approaches to a correct answer. The MC portion is machine scored and the CR and PE portions are hand-scored by hired raters who have been trained to read and score such items, using scoring guides (or rubrics).

The students may receive 0 or 1 point for a multiple-choice question; 0, 1, or 2 points for a constructed response question; and 0, 1, 2, 3, or 4 points on a performance

event. These are raw score points. The information about the weighting of individual questions that leads to the scale score is not made available to the public by the state.

DESE and CTB McGraw-Hill used the "bookmark procedure" to set the five achievement levels. These achievement levels are tied to scale score points.

> A panel composed of 40 to 45 teachers, parents, and business professionals reviewed the rank ordered test items from field-testing of the MAP. Test items were rank ordered from easiest to the most difficult based upon student performance during the field test. The panelists placed a bookmark at the point that they thought a student performing at Advanced, Proficient, Nearing Proficient, or Progressing would perform. The panelists then discussed the rationale for their judgments. The judgments of the panel members were averaged to establish cut off points for each achievement level. (Bratberg, 2002, p. 11)

The range of scale score points for each achievement level follows:

Grade 8 Mathematics:	Grade 10 Mathematics:
Advanced/Level 5 MAP score range: 785-915	Advanced/Level 5 MAP score range: 832-979
Proficient/Level 4 MAP score range: 744-784	Proficient/Level 4 MAP score range: 784-831
Nearing Proficient/Level 3 MAP score range: 708-743	Nearing Proficient/Level 3 MAP score range: 743-783
Progressing/Level 2 MAP score range: 668-707	Progressing/Level 2 MAP score range: 701-742
Step 1/Level 1 MAP score range: 541-667	Step 1/Level 1 MAP score range: 581-700

Procedures

The researcher obtained Human Subjects Research approval from the Institutional Review Board (IRB) at the University of Missouri St Louis. Data collection began upon approval of the research project by the IRB, the school district superintendent, the high school principal, the dissertation committee, and the graduate school.

Each individual student's name was matched with a unique student identification number. Only the researcher knows the student name and identification number matching. The student data was entered into the Statistical Package for the Social Sciences (SPSS) using the student identification number. The student data was analyzed and subjects were only included in the study if they had valid scores in each data field.

Data collection and analysis

Data included student test scores for Missouri Assessment Program (MAP) tests in mathematics for grade 8 and 10. The data collected included the following information for each student: for grades 8 and 10 MAP performance level score (1-5), raw score, scale score, scale score converted to Z-scores using the means and standard deviation from the state population for each test administration, Z-scores converted to T scores, points earned in each content area, and points earned on each item. The TerraNova scale score and percentile score that is part of the MAP 8 and 10 score report was also collected and converted to an NCE score. The ninth grade TerraNova scores (NCE) for Mathematics were included in the data. The eighth-grade and ninth-grade TerraNova Communication Arts (NCE) scores were also included in the data collected. Other data collected were student demographics (gender and race). Although the Missouri

Department of Elementary and Secondary Education recognizes six categories for race (Asian, Black, Hispanic, Native American or Alaskan Native, Pacific Islander, and White), the district that participated in this study only had students in sufficient numbers in the racial categories of Black and White. The few available scores for students listed as Asian, Hispanic, Native American or Alaskan Native, or Pacific Islander had scores that were similar to those of the students listed as White. Those students' scores were included with the students who were listed as White. The racial categories in this study were Black and White, with the White category representing all students who were not coded as Black. Transcript information that was collected included courses and grades. All grades were non-weighted and the grading scale used was A = 4.0, B+ = 3.33, B = 3.0, B- = 2.67, C+ = 2.33, C = 2.0, C- = 0.67, D+ = 1.33, D = 1.0, D- = 0.67, F = 0.

Design and Statistical Analysis

This research design is causal-comparative. Gall et al. (2003) advocate this type of design to study relationships when "experimental manipulation is difficult or impossible" (p. 298). Since this was an ex-post facto study using archival data, there was no possibility of an experimental procedure. In addition, it would be unethical to assign students to courses experimentally only to study the effects of course taking if those course placements were not known to be best for the students.

Researchers have found differential performance by gender on various item types (Anderson, 2002; Bielinski & Davison, 1998; Burton, 1996; Lane, Wang, & Magone, 1996; Metcalf, 2002; Muthen et al., 1995; Myerberg, 1996). The state does not report student performance by item type at the state, district, or building level. It is only by

conducting a local analysis of the data by item type that differential performance by gender on various item types can be detected. There is no published research of this type regarding performance on the MAP mathematics test.

Research Question One: What is the relationship between student gender and mean percentage correct for the three item types (multiple-choice, constructed response, performance event) on the grade eight or the grade 10 MAP mathematics test?

Hypothesis One: There is a statistically significant relationship between gender and item type on points received for items on the grade eight MAP mathematics test.

Hypothesis Two: There is a statistically significant relationship between gender and item type on points received for items on the grade 10 MAP mathematics test.

Research question one was analyzed using two-way repeated-measures analysis of variance. A separate analysis was conducted for grade eight and grade ten. The dependent variable was the mean of the percent of points received on each item type for each gender group at each grade level. The independent variables were the gender factor and the item type. The interaction between gender and item type was also analyzed. The data used were the percentage of total points available that were received by each student

on each item type at a grade level. From this data the mean of these percentages was computed for each gender on each item type.

Many researchers have found that student performance on questions within a content strand may be related to gender and/or course taking (Beller & Gafni, 1996; Bevan, 2001; Bielinski & Davison, 1998; Brosnan, 1998; Harris & Carlton, 1993; Metcalf, 2002). It is not known what effect a course in Algebra I in grade eight has on student performance in the six content strands on the MAP mathematics test. It is also not known what interaction effect can be observed between course taking and gender on the MAP content items in mathematics. Researchers have found that differential course taking outside of mathematics may contribute to performance on content strands (Metcalf, 2002). Since high school students have more diverse course-taking patterns than middle school students, to minimize the effect of non-math course-taking behavior this analysis was conducted on grade eight scores only.

Research Question Two: What is the relationship between student gender, student course-taking behavior, and mean percentage of points received for the six content strands (Number Sense, Geometry and Spatial Sense, Data Analysis and Probability, Patterns and Relationships, Mathematical Systems, and Discrete Mathematics) on the grade eight MAP mathematics test?

Hypothesis Three: There is a statistically significant relationship between gender and course-taking behavior on student MAP 8 mathematics content strands test scores.

Research question two was analyzed using a two-way ANOVA. The independent variables were the gender factor and the course-taking level factor. The interaction of gender and course taking was also analyzed. The dependent variable was the mean for each gender by course-taking group on each content strand. A separate analysis was conducted for each of the six content strands. Although students in grade eight only have two course-taking options (Algebra or Pre-Algebra) in grade eight mathematics in the school district that participated in the study, the course-taking groups were divided into three levels for this analysis. Initially students who would not go on to take Algebra in grade nine were included in the Pre-Algebra group. The scores for these students depressed the average achievement when compared with the students who would complete Algebra in grade nine to an extent that they distorted the situation. Therefore, these students scores were considered in a third group separated from those who took Algebra in grade eight or would take Algebra in grade nine. The three course-taking levels used in this analysis were:

Level 4: Algebra I in grade eight

Level 3: No Algebra I in grade eight but Algebra I in grade nine

Level 2+1: Algebra I or Algebra B in grade ten or Algebra I not completed by the end of grade 10.

Researchers have found that students who stay in the mathematics pipeline outperform those who drop out (Armstrong, 1981; Atanda, 2000; Burkam & Lee, 2003; Lee & Ware, 1986; Moses, Howe, & Niesz, 1999; Partenheimer & Miller, 2001; Rebhorn & Miles, 1999; Schiller & Muller, 2003; Sebring, 1985). There is also evidence in the research that students who take a more rigorous curriculum show greater growth in

achievement over time when measured by a norm-referenced test (Rock & Pollack, 1995b; Wang, 1999).

Student course taking is related to many factors outside the scope of this study: attitudes, parent aspirations for students, and aptitude. Student achievement in mathematics may be related to factors other than course-taking behavior. Researchers have questioned the content validity of mathematics constructed response and performance items, because of their reliance on reading and writing. Wilson and Zhang (1999) questioned whether an item that requires reading and writing skills is more a measure of mathematics achievement or communication arts achievement. Other researchers have stated that reading skills and mathematics skills go hand in hand. "Poor reading is highly correlated with poor mathematics achievement…Conversely, there is a strong correlation between high achievement in reading and in mathematics" (Czujko & Bernstein, 1989, p. 27). Some researchers indicate that measures of verbal skill indicate overall intelligence, "Verbal ability or achievement is a good index of intelligence" (Jones et al., 1986, p. 200). In order to identify factors related to MAP scores, the scores from the TerraNova Communication Arts and the MAP-8 and MAP 10 were analyzed by using a correlation matrix. The statistical method used was the Pearson product-moment correlation technique. Pearson correlation is appropriate for data that are continuous, as in the case of MAP T-scores and TerraNova NCE scores. The results of this correlation provided an r value indicating the strength of the relationship between reading/language scores and mathematics scores. If the strength of the relationship for reading or language with mathematics was significant at the 0.05 level, the MAP mathematics dependent variable was controlled for TerraNova reading/language scores in order to remove the

variance attributable to reading/language ability. Pearson is the preferred technique for such applications because of its small standard error (Gall et al., 1996). MAP T-scores and NCE scores for the subtests of TerraNova were used to calculate correlation coefficient, r. In a study of this type, researchers should "state and test hypotheses about other factors that might explain observed differences between two groups" (Gall et al., 1996).

Research Question Three: What is the relationship between gender, student course-taking behavior and Missouri Assessment Program (MAP) mathematics outcomes in grade eight and in grade 10?

Hypothesis Four: There is a statistically significant relationship between gender and course-taking behavior on the MAP 8 mathematics after taking into account the TerraNova grade 8 Language performance of students.

Hypothesis Five: There is a statistically significant relationship between gender and course-taking behavior on the MAP 10 mathematics after taking into account the TerraNova grade 9 Language performance of students.

The primary method of quantitative analysis for research question three was Analysis of Covariance (ANCOVA). The dependent variable was the MAP mathematics T-score for grade 8 or 10.

In order to isolate the variance attributable to mathematics from the potentially confounding variable of reading/language ability, the grade eight TerraNova language scores of the eighth grade subjects were used as the covariate in the ANCOVA. The

grade nine TerraNova language scores were used as the covariate for the 10th grade subjects. Because TerraNova scores were not available for eighth graders in the Student Cohort A (grade eight in 1998), this cohort was dropped from the grade eight analysis only. The sample for the Hypothesis four analysis included only students who were in grade eight in 1999, 2000, or 2001 and had valid scores in each data field.

The independent variables for the analyses were the gender factor and the course-taking behavior factor. The interaction between course taking and gender was also analyzed. There were four levels possible for course taking.

Level 4: Algebra I completed in grade eight.

Level 3: Algebra I completed in grade nine.

Level 2: Algebra I or Algebra B (the second part of a two year Algebra Series) completed in grade ten.

Level 1: Algebra I or Algebra B not completed by the end of 10^{th} grade.

These levels represent all of the possible options for completion of Algebra I in the district participating in the study. The accelerated group is represented by Level 4. This group of students is on track to take Advanced Placement (AP) Calculus in grade 12 if they remain in the mathematics pipeline. Researchers have found that these groups of students score significantly better on measures of mathematics achievement than non-accelerated groups (Gamoran, 1987; Jones et al., 1986; Kulik & Kulik, 1984; Ma, 2000; Smith, 1996). However, none of the research cited uses a standards-based, criterion-referenced test like the MAP as the dependent variable. It is not known how students in a course in Algebra in grade eight will perform relative to other student groups on a standards-based measure such as the MAP.

Level 3 represents students who begin high school at the level of the majority of students in the US, in an Algebra I class in ninth grade. These students are on track to take an advanced mathematics course such as College Algebra or Precalculus in grade 12 if they remain in the mathematics pipeline. Although there are only two levels of courses available in grade eight in the participating district (either Algebra or Pre-algebra), the students in the Pre-algebra group will be stratified into two different levels in grade nine. The district mathematics placement policy requires eighth-grade students to meet with mathematics teachers and high school counselors early in the second semester of grade eight, to determine their ninth-grade schedules. At that time they are sorted into Algebra in grade nine (Level 2) or Algebra A in grade nine (Level 3). Algebra A is the first course in a two year Algebra I sequence.

In *A Nation at Risk,* the National Council on Excellence in Education recommended a minimum common core for all students, the *Five New Basics* (National Commission on Excellence in Education, 1983). This recommendation included three years of mathematics in high school. ACT recommends a core curriculum in high school that includes at least three years of mathematics in high school, beginning with Algebra I. ACT reports that students who complete the ACT core have higher ACT scores and "are likely to fare better in college than those who don't" (www.act.org/news/releases/2003/8-20-03.html). The Level 2 students would not complete Algebra I until the successful completion of both Algebra A and Algebra B. If the Level 2 students are successful in the two-year Algebra sequence and they remain in the mathematics pipeline, they can still complete the ACT recommended core of three courses beginning with Algebra I by the end of grade 12. The Level 1 students have not completed Algebra I by the end of grade

10. Without doubling up or gaining credits in an alternative setting, summer school, or through correspondence, these students will be unable to complete the ACT recommended core. There is evidence that higher levels of mathematics coursework completed in high school correspond to greater success rates for students in post-secondary pursuits (Adelman, 1999; Bohr, 1994; Bottoms & Presson, 2000; Rose, 2001; Roth, Crans, & Carter, 2001; Schiller & Muller, 2003; U.S. Department of Education, 1997).

Research Question Four: Can the proficiency level(s) on the 10th grade Missouri Assessment Program (MAP) be predicted by some combination of the factors of mathematics course taking, performance on the eighth grade MAP mathematics test, TerraNova Comprehensive Tests of Basic Skills (CTBS) in Communication Arts in grade 9, student grade point average (GPA) in mathematics for grades 8 through 10, race, or gender?

Hypothesis Six: There is a statistically significant model using a combination of the factors of course-taking behavior, MAP-8 mathematics proficiency, TerraNova reading scores in grade nine, TerraNova language scores in grade nine, GPA in mathematics for grades 8 through 10, race, or gender, that predicts MAP 10 mathematics proficiency better than the constant-only model.

Research question three is a predictive analysis. The primary method of quantitative analysis was logistic regression in SPSS using the MAP-10 proficiency (0 =

not proficient, 1 = proficient) as the dependent variable. The seven independent variables considered for inclusion in the predictive model were gender (Female = 0, Male = 1), race (Black = 0, White = 1), course taking (Algebra in grade 8, no = 0, yes = 1), MAP-8 T-scores, grade-point average in mathematics for grades 8 through 10 (the average of available semester mathematics grades in grades 8 through 10. This was a number from 0 through 4), and scores on TerraNova Communication Arts in grade nine (one NCE score for reading and one NCE score for Language). The criterion variable was the MAP-10 score. It was considered a dichotomous variable with Levels 4 and 5 as "proficient" and Levels 1, 2, and 3 as "not proficient." The Missouri Department of Elementary and Secondary Education has stated that Level 4 (Proficient) is "the desired level for all students." In addition, the Missouri NCLB plan uses the MAP achievement level scores to identify the percent of students who are Proficient or above for the purposes of showing Adequate Yearly Progress. Level 4 on the MAP is also the required level for proficiency for NCLB accountability.

Research Question Five: Is there a relationship between scores on the TerraNova mathematics test in grades 8, 9, and 10 and course taking in mathematics?

Hypothesis Seven: There is a statistically significant relationship between course-taking behavior and NCE scores on the TerraNova test in grades 8, 9, and 10.

Research question five was analyzed using repeated measures ANOVA. The independent variables were time (grades 8, 9, or 10) and the course-taking factor described below. The data used for this analysis were the NCE scores from the

TerraNova portion of the MAP mathematics test in grades 8 and 10 and the NCE score for the mathematics portion of the TerraNova Multiple Assessments in grade 9. This analysis used the four course taking levels defined by:

Level 4: Algebra I completed in grade eight.

Level 3: Algebra I completed in grade nine.

Level 2: Algebra I or Algebra B (the second part of a two year Algebra Series) completed in grade ten.

Level 1: Algebra I or Algebra B not completed by the end of 10th grade.

The groups for this analysis were defined by course-taking level. The means of the group scale scores were analyzed for grades 8, 9, and 10 using repeated measures Analysis of Variance.

All data were analyzed using SPSS, Version 13 and all statistical tests of significance were at the .05 level.

Human Subjects Concerns

The data used were archival data held by the district. The data will only be used for this study. The data will be secured at my home for five years after the completion of the study. After five years, the data will be destroyed. Confidentiality will be maintained because only the primary researcher knows the names of the students. The data were coded for presentation in the dissertation. Permission has been obtained from the superintendent and the high school principal to conduct the study.

No harm is anticipated from participation in this study. The superintendent will be provided with a copy of the study results. The district can gain valuable information

about policy and practice related to course taking, guidance, and student achievement. The district will be identified only by the description (locale, demographics, population).

Chapter III outlined methods that were used in collecting and analyzing the data for this study, information about subjects, research design of the study, instruments, procedures, and human subjects concerns. The results of the statistical analyses will be reported in Chapter IV. Chapter V includes a summary of the study, discussion of the research findings, conclusions, implications of the findings, and recommendations for future research.

CHAPTER IV

Results

This study was a causal-comparative design. The data used in this study were examined using regression and correlation techniques to describe relationships and to determine if a model could be developed to predict performance on the MAP in grade 10. Analysis of variance was used to examine relationships between course-taking behavior, gender, and performance on the Missouri Assessment Program (MAP) mathematics tests in grades 8 and 10, and the TerraNova mathematics test in grade 9. Logistic regression was used to identify a model that would predict student proficiency on the MAP mathematics test in grade 10. The results of analyses proposed in Chapter 3 are summarized in this chapter. Each of the 7 hypotheses is listed, followed by the descriptive statistics tables, related figures, and a statement of the results for hypotheses tests. An alpha level of 0.05 was used for all statistical tests. SPSS, Version 13 was used for all analyses. The summary tables for the statistical analyses are reported in Appendix B.

In analysis of variance designs, certain assumptions must be satisfied in order to draw valid inferences from the data. These include:

1. normality of sampling distributions,
2. linearity,
3. homogeneity of variance
4. homogeneity of covariance, and
5. sphericity (in the case of repeated measures ANOVA).

Results of the evaluation of the assumptions of normality of sampling distributions and linearity were satisfactory. Concerning the issue of homogeneity of variance, Sherman (1989, p. 63) states, "this assumption can be satisfied even when the variance of the treatment effects for individual subjects differs considerably" by groups. She cited Lindquist:

> The assumption of homogeneity of variance is practically never strictly satisfied in educational and psychological experiments, but in most instances, the heterogeneity is not marked. Fortunately, the form of the sampling distribution of the mean square ratios is not very markedly affected by moderate degrees of heterogeneity of variance, and hence, the F-test may still be satisfactorily used in many experimental situations (Lindquist, 1953, p. 77-78).

Based on these statements, homogeneity of variance was assumed and not considered a critical issue. Similarly homogeneity of covariance was not considered a critical issue in the current research.

In repeated measures analyses where sphericity was violated, the Huynh-Feldt correction was applied. Although degrees of freedom and F-values were adjusted by this correction, the overall results related to the research hypotheses were not changed by application of the correction.

Hypothesis One (Two-way ANOVA with Repeated Measures)

The dependent variable for Hypothesis One (mean percentage correct for each item type on the grade 8 MAP mathematics test) is continuous in nature and the

independent variables (gender and item type) are categorical, making analysis of variance an appropriate method of analysis.

Hypothesis One: There is a statistically significant relationship between gender and item type on points received for items on the grade 8 MAP mathematics test.

Table 6 displays descriptive statistics for the dependent variable (mean percentage correct for each gender and item type on the grade 8 MAP mathematics test) disaggregated by independent variables (gender and item type). The cell sizes and cell size ratios are appropriate for ANOVA. The summary table for the repeated measures ANOVA is reported in Appendix Table B1.

Table 6

MAP mathematics grade 8 item type and gender, Group n, means, and standard Deviations

Item Type	*N*	*M*	*SD*
Constructed Response			
Females	264	49.3	23.0
Males	248	51.5	24.6
Total	512	50.4	23.8
Multiple Choice			
Females	264	70.8	16.2
Males	248	72.1	19.3
Total	512	71.4	17.8
Performance Event			
Females	264	39.9	27.6
Males	248	44.3	30.0
Total	512	42.0	28.9

Results of the evaluation of the assumptions of normality of sampling distributions, linearity, and homogeneity of variance were satisfactory. The assumption of sphericity was met after the Huynh-Feldt correction ($\varepsilon = 0.81$) was applied. Using Wilks' λ, the results indicated a significant within-subjects *item type* effect, Wilks' $\lambda = 0.29$, $F(2, 509) = 637.46$, $p < 0.001$. The *item type by gender* interaction was not significant, Wilks' $\lambda = 0.996$, $F(2, 509) = 1.099$, $p = 0.334$. The between-subjects main effect of *gender* also was not significant, $F(1, 510) = 2.008$, $p = 0.157$. The performance on the grade 8 MAP mathematics test for *gender by item type* is represented in Figure 1. The figure shows that males consistently performed better than females and that the profiles are the same. The figure also shows that both genders performed best on Multiple Choice, followed by Constructed Response, then Performance Events.

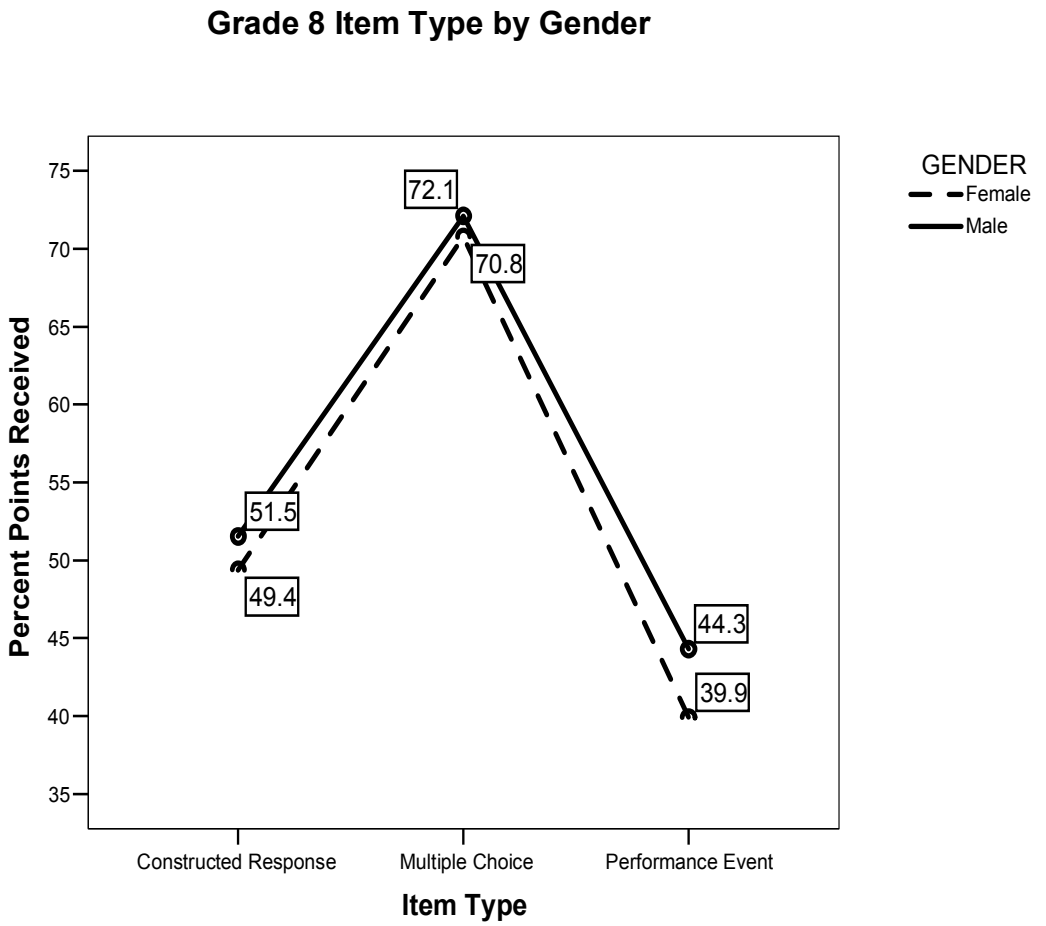

Figure 1. Percent of points received by each gender group for each item type in grade 8.

Hypothesis Two (Two-way ANOVA with repeated measures)

The dependent variable for Hypothesis Two (mean percentage correct for each item type on the grade 10 MAP mathematics test) is continuous in nature and the independent variables (gender and item type) are categorical, making analysis of variance an appropriate method of analysis.

Hypothesis Two: There is a statistically significant relationship between gender and item type on points received for items on the grade 10 MAP mathematics test.

Table 7 displays descriptive statistics for the dependent variable (mean percentage correct for each gender and item type on the grade 10 MAP mathematics test) disaggregated by independent variables (gender and item type). The cell sizes and cell size ratios are appropriate for ANOVA. The summary table for the repeated measures ANOVA is reported in Appendix Table B2.

Table 7

MAP mathematics grade 10 item type and gender, Group n, means, and standard deviations

Item Type	*N*	*M*	*SD*
Constructed Response			
Females	264	46.1	22.9
Males	248	49.8	25.7
Total	512	47.9	24.3
Multiple Choice			
Females	264	67.5	20.2
Males	248	71.9	21.7
Total	512	69.7	69.7
Performance Event			
Females	264	43.0	28.0
Males	248	47.1	31.7
Total	512	45.0	29.9

Results of the evaluation of the assumptions of normality of sampling distributions, linearity, and homogeneity of variance were satisfactory. The assumption of sphericity was met after the Huynh-Feldt correction ($\varepsilon = 0.86$) was applied. Using Wilks' λ, the results indicated a significant within-subjects *item type* effect,

Wilks' $\lambda = 0.25$, $F(2, 509) = 767.502$, $p < 0.001$. The *item type by gender* interaction was not significant Wilks' $\lambda = 0.999$, $F(2, 509) = 0.158$, $p = 0.854$. The results indicated a significant between-subjects main effect of *gender,* $F(1, 510) = 4.101$, $p = 0.043$. The performance on the grade 10 MAP mathematics test for *gender by item* type is represented in Figure 2. The figure shows that males consistently performed better than females and that the profiles are the same. The figure also shows that both genders performed best on Multiple Choice, followed by Constructed Response, then Performance Events.

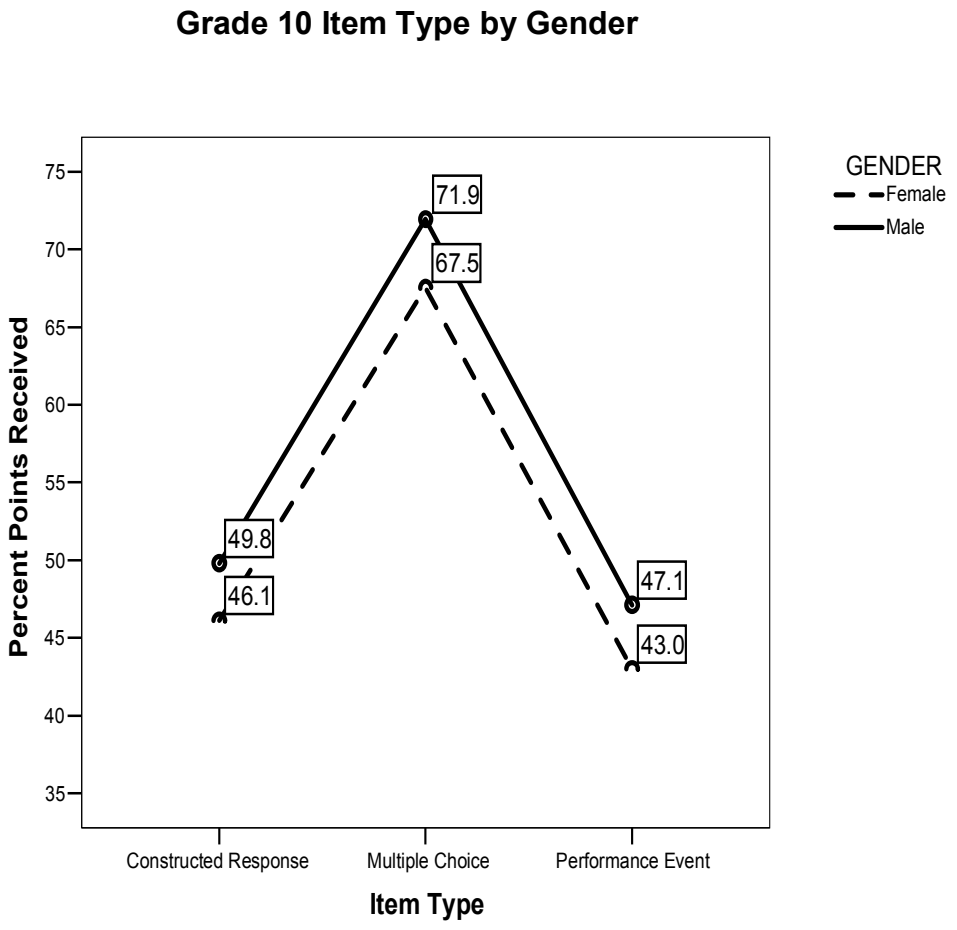

Figure 2. Percent of points received by each gender group for each item type in grade 10.

Hypothesis 3 (Two-way ANOVA)

The independent variables were the gender factor and course-taking level factor defined below. The interaction effect of gender by course taking was also analyzed. The dependent variable was the group means for the percent of points possible that were received by each group on each content strand. A separate analysis was completed for each of the six content strands. The six content strands are Number Sense, Geometry and Spatial Sense, Data Analysis and Probability, Patterns and Relationships, Mathematical Systems, and Discrete Mathematics.

The three course-taking levels that used for these analyses were:

Level 4: Algebra I in grade 8

Level 3: No Algebra I in grade 8 but Algebra I in grade 9

Level 2 + 1: Algebra I or Algebra B in grade 10 or Algebra I not completed by end of grade 10.

For this hypothesis only, Level 2 + 1 represents the combination of Levels 1 and 2 used for the other analyses. Although these students have not yet experienced different curriculum than the Level 3 students, their inclusion with Level 3 for these analyses masked performance on the various content strands by the Level 3 students.

The dependent variable for Hypothesis Three (mean percentage scores in each content strand in grade 8 mathematics) is continuous in nature, the independent variables (gender and course-taking behavior) are categorical making two-way analysis of variance an appropriate method of analysis.

Hypothesis Three: There is a statistically significant relationship between gender and course-taking behavior on student MAP 8 mathematics content strands test scores.

Table 8 presents descriptive statistics for the dependent variables (mean percentage scores in Number Sense on MAP mathematics in grade 8) disaggregated by independent variables (gender and course-taking category). The summary table for the two-way ANOVA is reported in Appendix Table B3. Figure 3 represents the performance on the MAP 8 Number Sense content strand for each gender by course-taking group.

Table 8

MAP mathematics grade 8 Number Sense content strand, gender, and course-taking, Group n, means, and standard deviations

Course Taking	*N*	*M*	*SD*
Algebra completed in grade 10 or Algebra not completed			
Females	51	43.4	17.6
Males	70	41.6	15.7
Total	121	42.3	16.5
Algebra completed in grade 9			
Females	127	60.6	14.9
Males	106	66.1	15.1
Total	233	63.1	15.2
Algebra completed in grade 8			
Females	86	81.5	12.9
Males	72	86.5	10.4
Total	158	83.8	12.1
Total			
Females	264	64.1	20.2
Males	248	65.1	22.1
Total	512	64.6	21.1

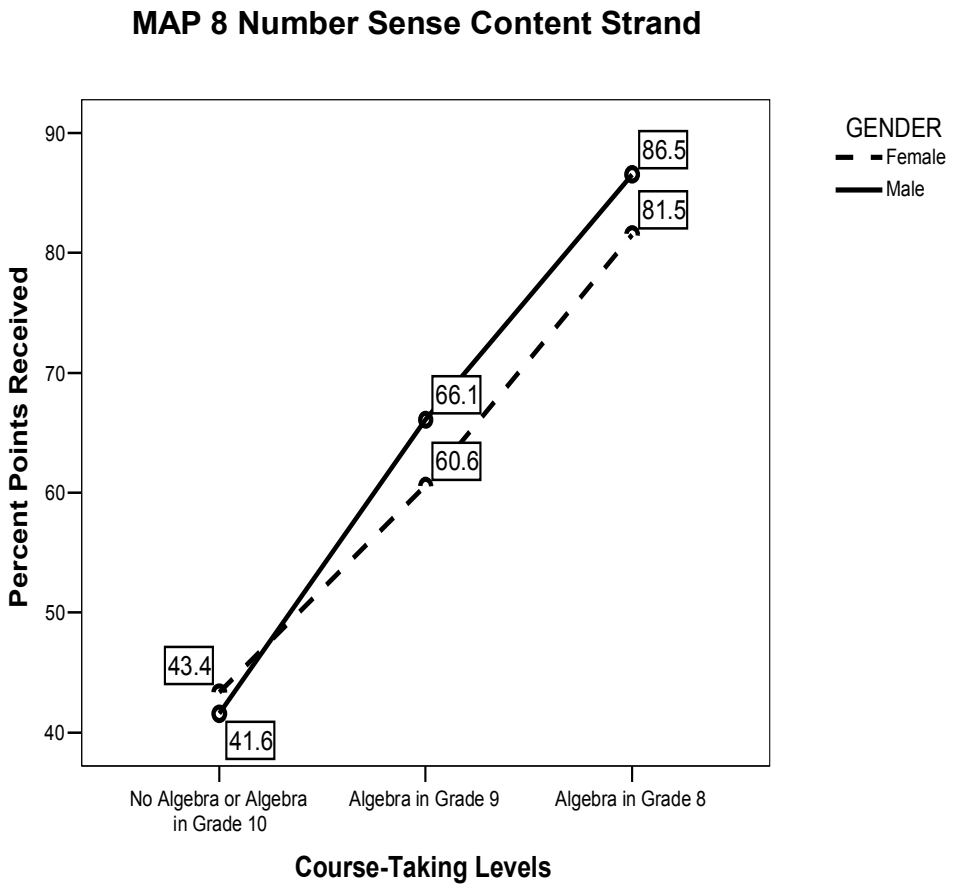

Figure 3. Percent of points received by each gender by course-taking group for the Number Sense content strand in grade 8.

Results of the evaluation of the assumptions of normality of sampling distributions, linearity, and homogeneity of variance were satisfactory. The results for the Number Sense content strand indicated a significant main effect of *gender*, $F(1, 506) = 4.770$, $p = 0.029$ and a significant main effect of *course-taking*, $F(2, 506) = 278.868$, $p < 0.001$. The *course-taking by gender* interaction was not significant, $F(2, 509) = 2.729$, $p = 0.854$. Figure 3 shows the significant effect of course taking with students who complete Algebra courses earlier earning higher scores on Number Sense. Although Figure 3 indicates an interaction between males and females, the scores at the lowest level of course taking were higher for females, this interaction was not statistically

significant. In the other two groups, the mean scores for males was higher than that for females.

Table 9 displays descriptive statistics for the dependent variables (mean percentage scores in Geometry and Spatial Sense on MAP mathematics in grade 8) disaggregated by independent variables (gender and course-taking category). The summary table for the two-way ANOVA is reported in Appendix Table B4. Figure 4 represents the performance on the grade 8 Geometry and Spatial Sense content strand for each gender by course-taking group.

Table 9

MAP mathematics grade 8 Geometry and Spatial Sense content strand, gender, and course-taking, Group n, means, and standard deviations

Course Taking	*N*	*M*	*SD*
Algebra completed in grade 10 or Algebra not completed			
Females	51	28.1	11.8
Males	70	25.3	14.5
Total	121	26.5	13.4
Algebra completed in grade 9			
Females	127	40.0	17.3
Males	106	47.8	19.6
Total	233	43.6	18.7
Algebra completed in grade 8			
Females	86	63.9	19.3
Males	72	78.4	14.7
Total	158	70.5	18.8
Total			
Females	264	45.5	21.8
Males	248	50.3	26.3
Total	512	47.8	24.2

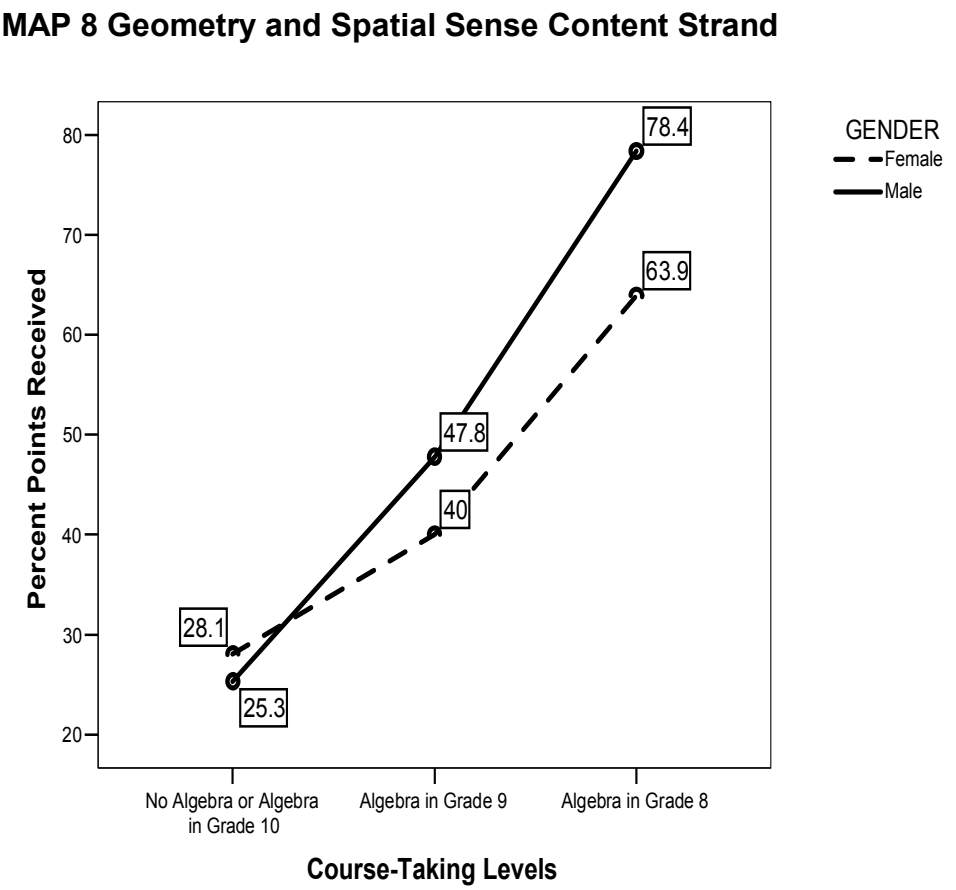

Figure 4. Percent of points received by each gender by course-taking group for the Geometry and Spatial Sense content strand in grade 8.

The results for the Geometry and Spatial Sense content strand indicated a significant interaction effect of *course-taking by gender*, $F(2, 509) = 8.633$, $p < 0.001$. Figure 4 shows the significant effect of course taking, with students who complete Algebra courses earlier earning higher scores on Geometry and Spatial Sense. Figure 4 indicates an interaction between males and females, the scores at the lowest level of course taking were higher for females, in the other two course-taking groups, the mean scores for males was higher than that for females. The difference between the genders favored females by 2.8 points, then favored males by 14.5 points at the highest course-taking level, representing a total gain of 17.3 points in mean difference for males.

Table 10 displays descriptive statistics for the dependent variables (mean percentage scores in Data Analysis and Probability on MAP mathematics in grade 8) disaggregated by independent variables (gender and course-taking category). The summary table for the two-way ANOVA is reported in Appendix Table B5. Figure 5 represents the performance on the grade 8 Data Analysis and Probability content strand for each gender by course-taking group.

Table 10

MAP mathematics grade 8 Data Analysis and Probability content strand, gender, and course-taking, Group n, means, and standard deviations

Course Taking	*N*	*M*	*SD*
Algebra completed in grade 10 or Algebra not completed			
Females	51	42.4	16.7
Males	70	40.4	19.1
Total	121	41.3	18.1
Algebra completed in grade 9			
Females	127	59.4	16.0
Males	106	64.2	14.6
Total	233	61.6	15.5
Algebra completed in grade 8			
Females	86	76.4	15.6
Males	72	81.8	13.4
Total	158	78.8	14.8
Total			
Females	264	61.6	20.0
Males	248	62.6	22.2
Total	512	62.1	21.0

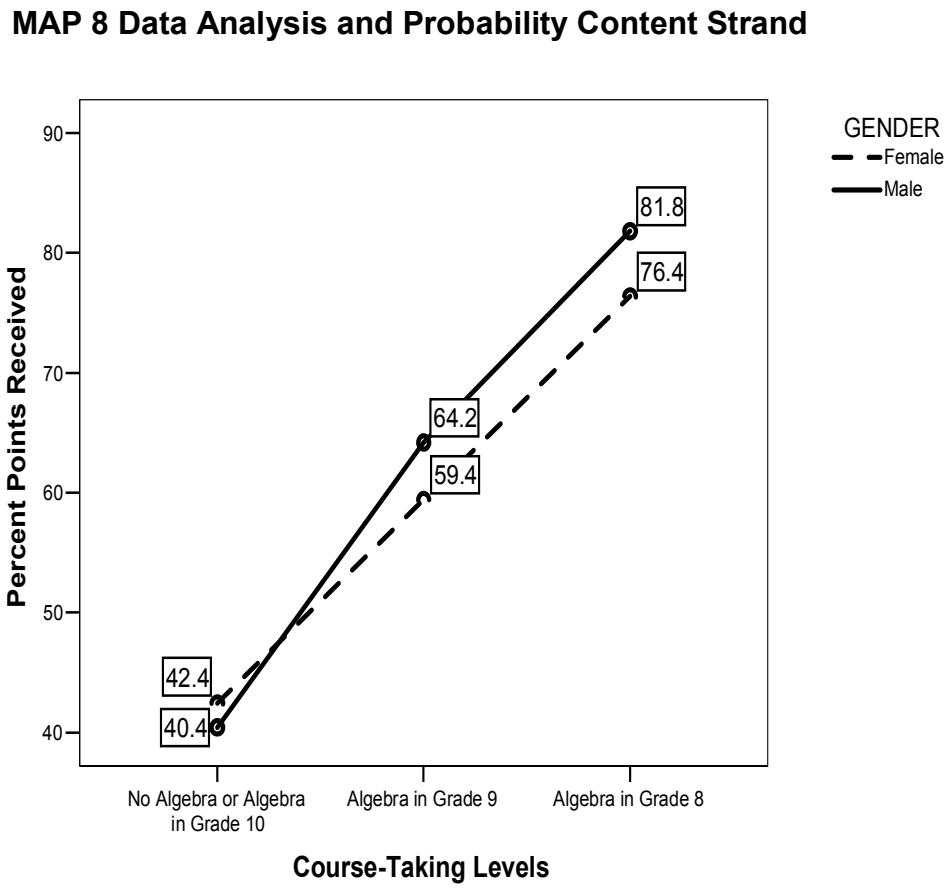

Figure 5. Percent of points received by each gender by course-taking group for the Data Analysis and Probability content strand in grade 8.

The results for the Data Analysis and Probability content strand indicated a significant main effect of *course-taking*, $F(2, 506) = 190.714$, $p < 0.001$.The results indicated no statistically significant main effect of *gender*, $F(1, 506) = 3.520$, $p < 0.061$. The *course-taking by gender* interaction was also not found to be statistically significant, $F(2, 509) = 2.266$, $p = 0.105$. Figure 5 illustrates the significant effect of course taking with students who complete Algebra courses earlier earning higher scores on Data Analysis and Probability. Although Figure 5 indicates an interaction between males and females, the scores at the lowest level of course taking were higher for females, this interaction was not statistically significant. In the other two course-taking groups, the

mean scores for males were higher than that for females; however, the gender factor was not a statistically significant main effect.

Table 11 displays descriptive statistics for the dependent variables (mean percentage scores in Patterns and Relationships on MAP mathematics in grade 8) disaggregated by independent variables (gender and course-taking category). The summary table for the two-way ANOVA is reported in Appendix Table B6. Figure 6 represents the performance on the grade 8 Patterns and Relationships content strand for each gender by course-taking group.

Table 11

MAP mathematics grade 8 Patterns and Relationships content strand, gender, and course-taking, Group n, means, and standard deviations

Course Taking	*N*	*M*	*SD*
Algebra completed in grade 10 or Algebra not completed			
Females	51	36.8	27.5
Males	70	33.3	24.4
Total	121	34.8	25.7
Algebra completed in grade 9			
Females	127	53.2	27.9
Males	106	60.2	27.2
Total	233	56.4	27.7
Algebra completed in grade 8			
Females	86	80.8	22.7
Males	72	84.3	16.5
Total	158	82.4	20.1
Total			
Females	264	59.0	30.8
Males	248	59.6	30.5
Total	512	59.3	30.7

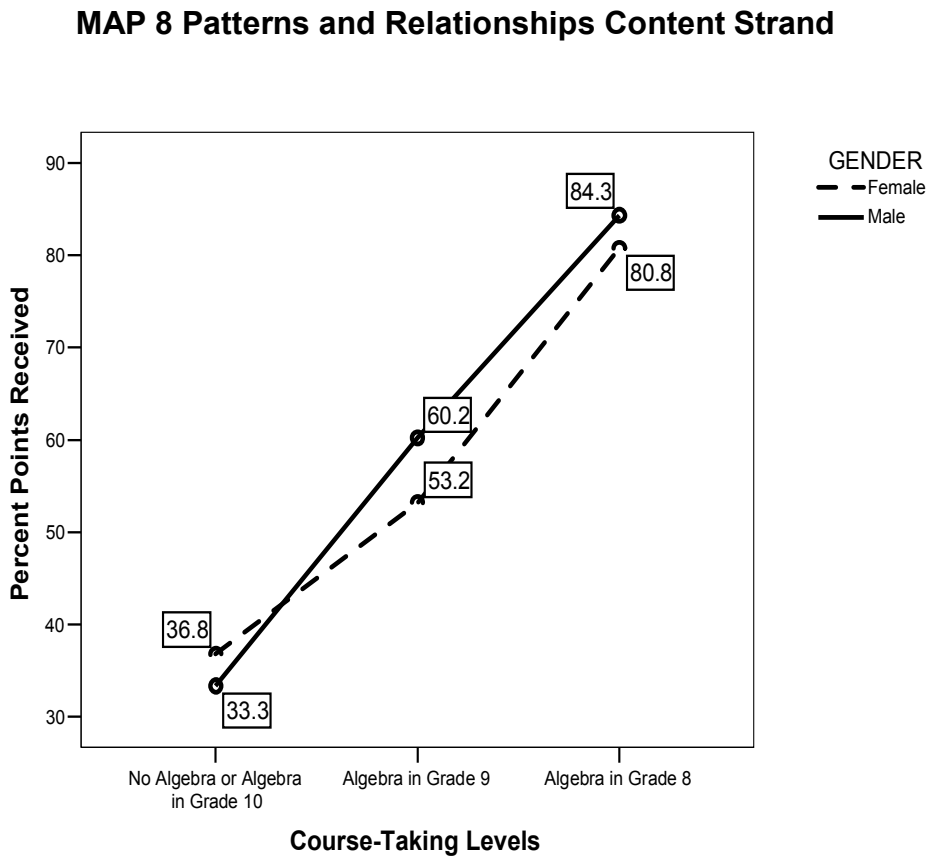

Figure 6. Percent of points received by each gender by course-taking group for the Patterns and Relationships content strand in grade 8.

The results for the Patterns and Relationships content strand indicated a significant main effect of *course-taking*, $F(2, 506) = 123.649$, $p < 0.001$. The results indicated no statistically significant main effect of *gender*, $F(1, 506) = 1.046$, $p = 0.307$. The *course-taking by gender* interaction was also not found to be statistically significant, $F(2, 509) = 1.705$, $p = 0.183$. Figure 6 shows the significant effect of course taking with students who complete Algebra courses earlier earning higher scores on Patterns and Relationships. Although Figure 6 indicates an interaction between males and females, the scores at the lowest level of course taking were higher for females, this interaction was not statistically significant. In the other two course-taking levels, the

mean scores for males was higher than that for females with the gap narrowing in the highest course-taking level. In this case, the gender factor was not a statistically significant main effect.

Table 12 displays descriptive statistics for the dependent variables (mean percentage scores in Mathematical Systems on MAP mathematics in grade 8) disaggregated by independent variables (gender and course-taking category). The summary table for the two-way ANOVA is reported in Appendix Table B7. Figure 7 represents the performance on the grade 8 Mathematical Systems content strand for each gender by course-taking group.

Table 12

MAP mathematics grade 8 Mathematical Systems content strand, gender, and course-taking, Group n, means, and standard deviations

Course Taking	*N*	*M*	*SD*
Algebra completed in grade 10 or Algebra not completed			
Females	51	37.1	19.2
Males	70	44.7	22.3
Total	121	41.5	21.3
Algebra completed in grade 9			
Females	127	63.6	21.1
Males	106	68.7	19.4
Total	233	65.9	20.5
Algebra completed in grade 8			
Females	86	77.1	19.8
Males	72	81.0	18.1
Total	158	78.9	19.1
Total			
Females	264	62.8	24.6
Males	248	65.5	24.3
Total	512	64.1	24.5

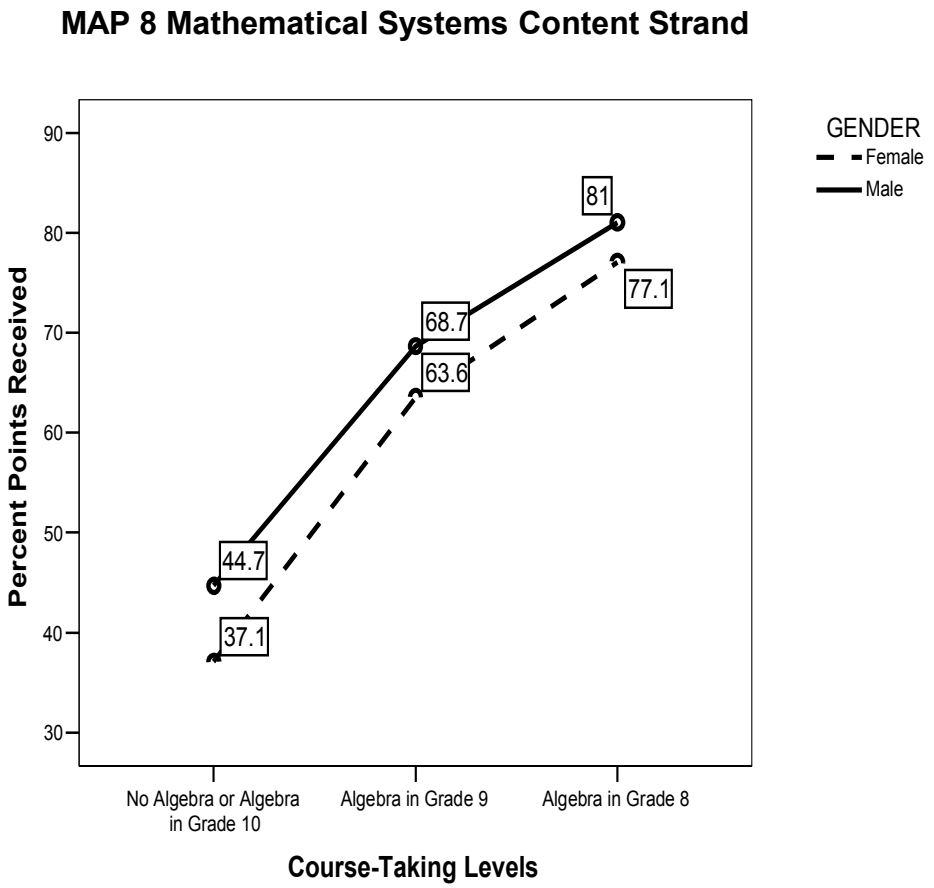

Figure 7. Percent of points received by each gender by course-taking group for the Mathematical Systems content strand in grade 8.

The results for the Mathematical Systems content strand indicated a significant main effect of *course-taking*, $F(2, 506) = 122.822$, $p < 0.001$ and a significant main effect of *gender*, $F(1, 506) = 8.978$, $p = 0.003$. The *course-taking by gender* interaction was not found to be statistically significant, $F(2, 509) = 0.283$, $p = 0.754$. Figure 7 shows the significant effect of course taking with students who complete Algebra courses earlier earning higher scores on Mathematical Systems. Figure 7 shows that the gender profiles were the same, with males consistently outscoring females in Mathematical Systems at all course-taking levels.

Table 13 displays descriptive statistics for the dependent variables (mean percentage scores in Discrete Mathematics on the MAP mathematics test in grade 8) disaggregated by independent variables (gender and course-taking category). The summary table for the two-way ANOVA is reported in Appendix Table B8. Figure 8 represents the performance on the grade 8 Discrete Mathematics content strand for each gender by course-taking group.

Table 13

MAP mathematics grade 8 Discrete Mathematics content strand, gender, and course-taking, Group n, means, and standard deviations

Course Taking	*N*	*M*	*SD*
Algebra completed in grade 10 or Algebra not completed			
Females	51	27.8	25.5
Males	70	23.5	26.4
Total	121	25.3	26.0
Algebra completed in grade 9			
Females	127	45.8	25.5
Males	106	51.2	26.3
Total	233	48.3	26.0
Algebra completed in grade 8			
Females	86	74.4	24.5
Males	72	74.7	24.6
Total	158	74.5	24.5
Total			
Females	264	51.6	30.4
Males	248	50.2	32.3
Total	512	50.9	31.3

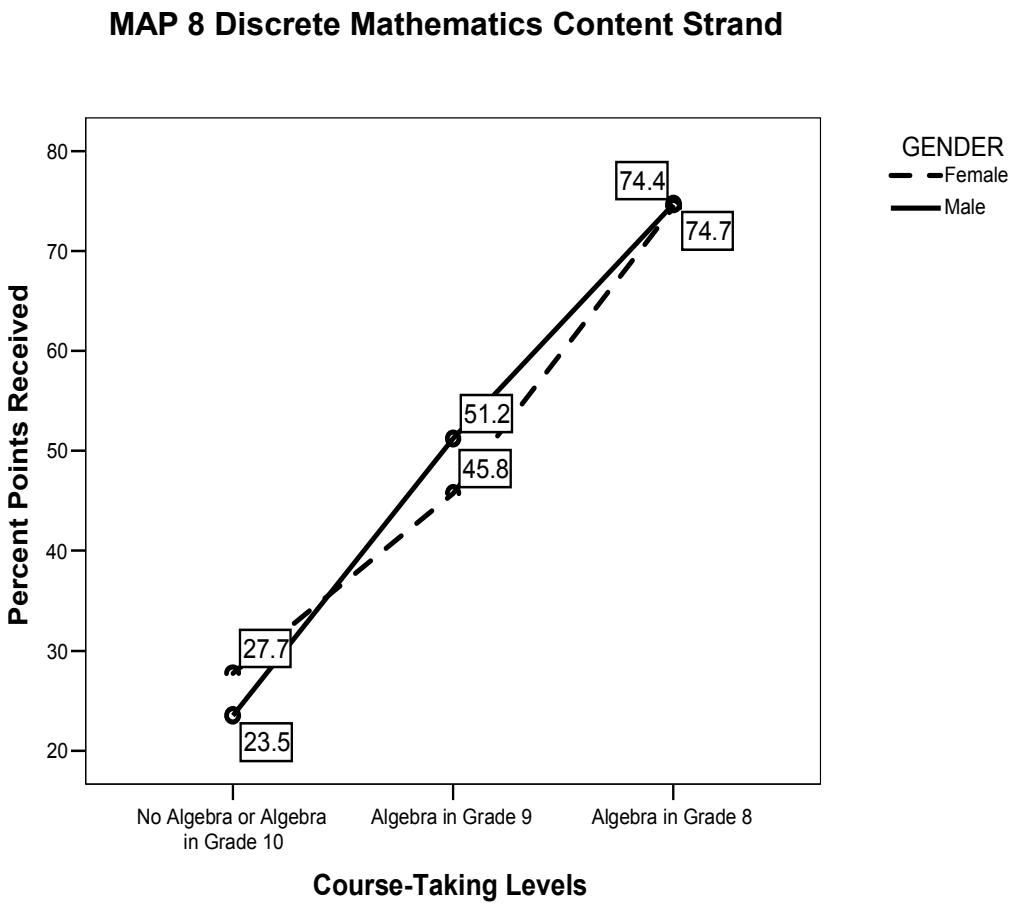

Figure 8. Percent of points received by each gender by course-taking group for the Discrete Mathematics content strand in grade 8.

The results for the Discrete Mathematics content strand indicated a significant main effect of *course-taking*, $F(2, 506) = 126.002$, $p < 0.001$.The results indicated no statistically significant main effect of *gender*, $F(1, 506) = 0.049$, $p = 0.824$. The *course-taking by gender* interaction was also not found to be statistically significant, $F(2, 509) = 1.484$, $p = 0.228$. Figure 8 shows scores increased as course-taking levels increased with a slight interaction at the lowest course-taking levels where females scored higher than males. At the highest course-taking levels, the scores for each gender were almost the same.

In summary, the analyses for Hypothesis Three indicated only the Geometry and Spatial Sense content strand had a statistically significant interaction effect of course-taking by gender. Course taking was a statistically significant main effect for each of the other five content strands. Gender was found to be a statistically significant main effect for two content strands: Number Sense and Mathematical Systems.

Hypothesis 4 (Two-way ANCOVA)

The independent variables were gender factor and course-taking level factor. The interaction effect of gender by course taking was also analyzed. The dependent variable was the mean for the scale scores (converted to T-scores) for each group on the MAP 8 mathematics test. The Missouri Department of Elementary and Secondary Education (DESE) provided the population means and standard deviations for each test administration included in the study. These population data were used to convert the subjects' scale scores to Z-scores. The Z-scores were then converted to T-scores. The four course-taking levels that were used for this analysis were:

Level 4: Algebra I in grade 8

Level 3: Algebra I in grade 9

Level 2: Algebra I or Algebra B in grade 10

Level 1: Algebra I not completed by the end of grade 10

The dependent variable for Hypothesis Four (mean T-scores on the MAP 8 mathematics test) was continuous in nature, the independent variables (gender and course-taking behavior) were categorical making analysis of covariance an appropriate method of analysis. The selected covariate was subjects' grade 8 Terra Nova Language

NCE score for the MAP 8 analysis. For this analysis, the sample was reduced to include only those students who were in grade 8 in 1999, 2000, and 2001. The 1998 cohort could not be included because they did not take the TerraNova test so they did not have valid scores for the covariate of TerraNova Language in grade 8.

ANCOVA focuses on detecting differences between groups "controlling" for the influence of extraneous variables that might otherwise confound the analysis. Using the covariate as a control allows better focus on the independent variables being analyzed. Ideally, the covariate should have a moderate to high correlation with the dependent variable. The correlation should not be so high as to be collinear.

Table 14 displays the Pearson r correlation between the covariate and the DV for ANCOVA. The correlation is moderate, at 0.737, and of an appropriate level for use as a covariate.

Table 14

Pearson correlation between MAP mathematics grade 8 scale score (DV) and Terra Nova grade 8 language scale score (Covariate)

Dependent variable	Statistic	Terra Nova grade 8 Language NCE (covariate)
MAP 8 T-score	Pearson r	.737**
	Sig. (2-tailed)	.000
	N	379

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

Hypothesis Four: There is a statistically significant relationship between gender and course-taking behavior on the MAP 8 mathematics after taking into account the TerraNova grade 8 Language performance of students.

Table 15 displays descriptive statistics for the dependent variable (mean T-scores on the MAP mathematics test in grade 8 after TerraNova grade 8 Language scores were used as the covariate) disaggregated by the independent variables (gender and course-taking category). The summary table for the two-way ANCOVA is reported in Appendix Table B9. Figure 9 shows the performance by course-taking levels for each gender on the MAP 8 Mathematics test, after the TerraNova grade 8 language scores were used as the covariate.

Table 15

MAP mathematics grade 8 MAP T-scores with TerraNova Language grade 8 covariate, gender, and course-taking, Group n, means, and standard error

Course Taking by Gender with TerraNova Language covariate	*N*	*M*	**SE**
Algebra not completed			
Females	13	46.1	1.5
Males	21	44.8	1.2
Total	34	45.4	1.0
Algebra completed in grade 10			
Females	22	44.1	1.2
Males	23	48.8	1.2
Total	45	46.5	0.9
Algebra completed in grade 9			
Females	100	51.4	0.5
Males	76	54.8	0.6
Total	176	53.1	0.4
Algebra completed in grade 8			
Females	68	56.4	0.7
Males	56	60.7	0.8
Total	124	58.5	0.6
Total			
Females	203	49.5	0.5
Males	176	52.3	0.5
Total	379	50.9	0.4

Note. MAP 8 scale scores were converted to Z-scores and then T scores, using population means and standard deviations provided by the Missouri Department of Elementary and Secondary Education.

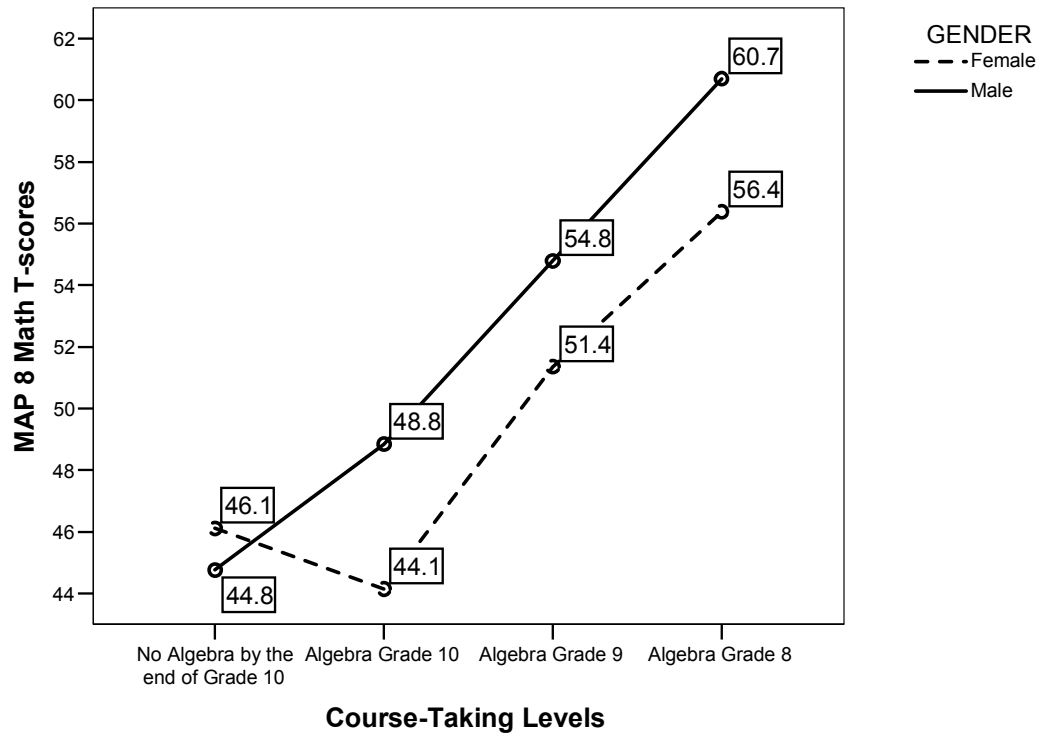

Figure 9. Group means for each gender by course-taking group on grade 8 MAP Mathematics T-scores, after using grade 8 TerraNova Language scores as the covariate.

Results of the evaluation of the assumptions of normality of sampling distributions, linearity, and homogeneity of variance were satisfactory. The results indicated a significant main effect of *gender*, $F(1, 370) = 15.372$, $p < 0.001$ and a significant main effect of *course-taking*, $F(3, 370) = 47.250$, $p < 0.001$. The *course-taking by gender* interaction was not statistically significant, $F(3, 370) = 2.588$, $p = 0.053$. The adjusted R Squared shows that this model explains approximately 70% of the variance in mean grade 8 MAP math T-scores.

Pairwise comparisons for course-taking levels showed significant differences between all pairs except course-taking levels 1 and 2. Figure 9 demonstrates that

performance on MAP 8 increases significantly for course-taking groups, over and above the effect of TerraNova Language scores. At the lowest course-taking level, females outscored males, at all other course-taking levels the gap between the genders favored males.

Hypothesis 5 (Two-way ANCOVA)

The independent variables were the gender factor and course-taking level factor. The interaction effect of gender by course taking was also analyzed. The dependent variable was the mean for the scale scores (converted to T-scores) for each group on the MAP 10 mathematics test. The Missouri Department of Elementary and Secondary Education (DESE) provided the population means and standard deviations for each test administration included in the study. These population data were used to convert the subjects' scale scores to Z-scores. The Z-scores were then converted to T-scores. The four course-taking levels used for this analysis were:

Level 4: Algebra I in grade 8

Level 3: Algebra I in grade 9

Level 2: Algebra I or Algebra B in grade 10

Level 1: Algebra I not completed by the end of grade 10

The dependent variable (mean T-scores on the MAP 10 mathematics test) for Hypothesis Five was continuous in nature, the independent variables (gender and course-taking behavior) were categorical, making analysis of covariance an appropriate method of analysis. The selected covariate was subjects' grade 9 Terra Nova Language NCE (or Reading) score for the MAP 10 analysis.

ANCOVA focuses on detecting differences between groups "controlling" for the influence of extraneous variables that might otherwise confound the analysis. Using the covariate as a control allows better focus on the independent variables being analyzed. Ideally, the covariate should have a moderate to high correlation with the dependent variable. The correlation should not be so high as to be collinear.

Table 16 displays the Pearson r correlation between the covariate and the DV for ANCOVA. The correlation is moderate, at 0.585, and of an appropriate level for use as a covariate.

Table 16

Pearson correlation between MAP mathematics grade 10 scale score (DV) and Terra Nova grade 9 language scale score (Covariate)

Dependent variable	Statistic	Terra Nova grade 9 Language NCE (covariate)
MAP 10 scale score	Pearson r	.585**
	Sig. (2-tailed)	.000
	N	512

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

Hypothesis Five: There is a statistically significant relationship between gender and course-taking behavior on the MAP 10 mathematics after taking into account the TerraNova grade 9 Language performance of students.

Table 17 displays descriptive statistics for the dependent variable (mean T-scores on the MAP mathematics test in grade 10 after TerraNova grade 9 Language scores were used as the covariate) disaggregated by the independent variables (gender and course-taking category). The summary table for the two-way ANCOVA is reported in Appendix Table B10. Figure 10 shows the performance by course-taking levels for each gender on

the MAP 10 Mathematics test, after the TerraNova grade 9 language scores were used as the covariate.

TABLE 17
MAP mathematics grade 10 MAP T-scores with TerraNova Language grade 9 covariate, gender, and course-taking, Group n, means, and standard error

Course Taking by Gender with TerraNova Language covariate	*N*	*M*	**SE**
Algebra not completed			
Females	21	44.0	1.5
Males	36	46.7	1.2
Total	57	45.3	1.0
Algebra completed in grade 10			
Females	31	44.5	1.3
Males	33	50.9	1.3
Total	64	47.7	0.9
Algebra completed in grade 9			
Females	127	49.8	0.6
Males	106	52.9	0.7
Total	233	51.4	0.5
Algebra completed in grade 8			
Females	86	57.1	0.8
Males	72	61.9	0.9
Total	158	59.5	0.7
Total			
Females	264	48.9	0.5
Males	248	53.1	0.5
Total	512	51.0	0.4

Note. MAP 10 scale scores were converted to Z-scores and then T scores, using population means and standard deviations provided by the Missouri Department of Elementary and Secondary Education.

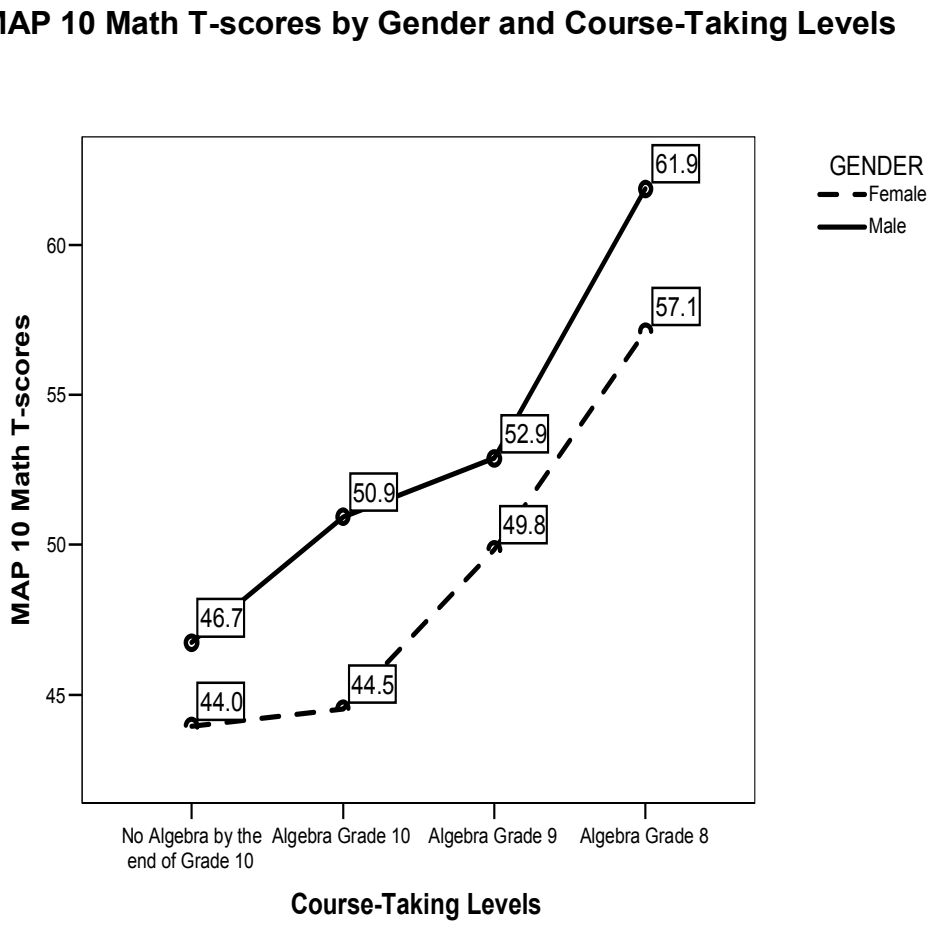

Figure 10. Group means for each gender by course-taking group on grade 8 MAP Mathematics T-scores, after using grade 8 TerraNova Language scores as the covariate.

Results of the evaluation of the assumptions of normality of sampling distributions, linearity, and homogeneity of variance were satisfactory. The results indicated a significant main effect of *gender*, $F(1, 503) = 31.590$, $p < 0.001$ and a significant main effect of *course-taking*, $F(3, 503) = 48.057$, $p < 0.001$. The *course-taking by gender* interaction was not statistically significant, $F(3, 503) = 1.311$, $p = 0.270$.

The adjusted R Squared shows that this model explains approximately 53% of the variance in mean grade 10 MAP math T scores.

Pairwise comparisons for course-taking levels revealed significant differences between all pairs except levels 1 and 2 and levels 2 and 3. Figure 10 illustrates that performance on MAP 10 increases for course-taking groups, even after using TerraNova Language scores as a covariate. Males' mean scores were higher than females for all course-taking levels.

Hypothesis 6 (Logistic Regression)

The dependent variable for Hypothesis Six (proficient/not proficient on MAP grade 10 mathematics) is dichotomous in nature and the independent variables are both categorical (course-taking behavior, race, and gender) and continuous (MAP-8 T-scores on the mathematics test, GPA in mathematics for grades 8 through 10, TerraNova grade 9 reading NCE scores, TerraNova grade 9 language NCE scores) making Logistic Regression an appropriate method of analysis. The purpose of this hypothesis is to determine whether there is a model that predicts MAP 10 achievement better than the model with only the constant and none of the predictor variables. The hypothesis for this logistic regression is:

> Hypothesis Six: There is a statistically significant model using a combination of the factors of course-taking behavior, MAP-8 mathematics proficiency, Terra Nova reading scores in grade 9, TerraNova language scores in grade 9, GPA in mathematics for grades 8 through 10, race, or gender, that predicts MAP 10 mathematics proficiency better than the constant-only model.

The objective of Logistic Regression Analysis is to find a model that significantly improves on the constant-only model using the least number of predictor variables. The

constant only model predicts all cases as Not Proficient by default. This constant-only default prediction led to a successful prediction rate of 81.8%. Using the SPSS "Enter" method for Logistic Regression, the best model found included the following predictors: MAP 8 T-scores, Math GPA for grades 8 through 10 (the mean of semester math grades in grades 8 through 10), Course-Taking Level 4 (1 = taking Algebra in grade 8 and 0 = not taking Algebra I grade 8), and gender (1 = Male, 0 = Female). This model improved the prediction rate to 91.4%. The prediction rate for the outcome of interest, MAP 10 Proficiency, improved from 0% to 67.7%. The model including these predictors improved over a constant only model, $\chi^2(4, 512) = 244.649$, $p < 0.001$, with goodness of fit , $\chi^2(8, 512) = 17.888$, $p = 0.022$.

The equation for the model is represented by:

$$\ln (p/(1-p)) = -15.325 + 0.153 X_1 + 1.238 X_2 + 1.513 X_3 + 1.013 X_4$$

ln (p/(1-p)) represents the logit or the natural logarithm of the ratio of the probability of scoring Proficient in grade 10 to the probability of not scoring Proficient in grade 10.

X_1 represents MAP 8 T-scores

X_2 represents MATH GPA for semesters in grades 8 through 10

X_3 represents Course-Taking Level 4 (Algebra in grade 8); 1 = yes, 0 = no

X_4 represents Gender; 1 = male, 0 = female

Evaluation of individual predictors indicated MAP 8 (Wald = 17.128, Odds Ratio = 1.165, CI = 1.084-1.252), Math GPA (Wald = 17.482, Odds Ratio = 3.450, CI =

1.931-6.166), Course Taking Level 4 (Wald = 13.303, Odds Ratio = 4.541, CI = 2.014-10.239), and gender (Wald = 8.437, Odds Ratio = 2.753, CI = 1.390-5.452) were all significant predictors.

The coefficients in logistic regression are expressed in terms of the natural log of the odds (http://www.ats.ucla.edu/stat/stata/faq/oratio.htm). In the case of MATH GPA, the coefficient B indicates that a one-unit change in Math GPA would result in a 1.238 change in the log of the odds. Similar changes can be interpreted for the other three variables. The summary table for the logistic regression is reported in Appendix Table B11.

Hypothesis 7 (Repeated measures ANOVA)

The dependent variable for Hypothesis Seven (mean NCE scores for Terra Nova Mathematics test in grades 8, 9, and 10) is continuous in nature and the independent variable (course-taking behavior) is categorical, making analysis of variance an appropriate method of analysis. The independent variables are grade level (grades 8, 9, or 10) and the course-taking factor described below. The data used for this analysis were the NCE scores from the Terra Nova portion of the MAP mathematics test in grades 8 and 10 and the NCE score for the mathematics portion of the Terra Nova Multiple Assessments in grade 9. This analysis used the four course taking levels defined by:

Level 4: Algebra I completed in grade 8.

Level 3: Algebra I completed in grade 9.

Level 2: Algebra I or Algebra B (the second part of a two year Algebra Series) completed in grade 10.

Level 1: Algebra I or Algebra B not completed by the end of 10th grade.

The means of the NCE scores on the TerraNova for the four course-taking groups were analyzed for grades 8, 9, and 10 using repeated measures Analysis of Variance.

Hypothesis Seven: There is a statistically significant relationship between course-taking behavior and NCE scores on the TerraNova test in grades 8, 9, and 10.

Table 18 displays descriptive statistics for the dependent variable (mean NCE mathematics scores for the TerraNova) disaggregated by independent variables (grade level and course taking). The cell sizes and cell size ratios are appropriate for ANOVA. The summary table for the repeated measures ANOVA is reported in Appendix Table B12. Figure 11 shows the group mean Mathematics NCE scores for TerraNova mathematics tests for each course-taking group at grades 8, 9, and 10.

Table 18

TerraNova mathematics NCE scores for grades 8, 9, and 10 by course-taking levels, Group n, means, and standard deviations

Course Taking	*M*	**SD**
Algebra not completed (n = 57)		
Grade Level 8	38.9	17.2
Grade Level 9	36.4	13.0
Grade Level 10	45.6	18.9
Algebra completed in grade 10 (n = 64)		
Grade Level 8	42.4	15.2
Grade Level 9	41.4	14.5
Grade Level 10	50.7	20.5
Algebra completed in grade 9 (n = 233)		
Grade Level 8	60.4	13.3
Grade Level 9	56.3	12.9
Grade Level 10	62.5	16.2
Algebra completed in grade 8 (n = 158)		
Grade Level 8	81.1	12.7
Grade Level 9	76.8	13.0
Grade Level 10	84.6	13.3
Totals (n = 512)		
Grade Level 8	62.1	20.3
Grade Level 9	58.6	19.2
Grade Level 10	65.9	21.3

Note: The TerraNova portion of the MAP in grades 8 and 10 is the TerraNova Survey. The TerraNova in grade 9 is the mathematics portion of the TerraNova Multiple Assessments. All TerraNova formats are scored on a common scale and NCEs from different forms can be compared from year to year.

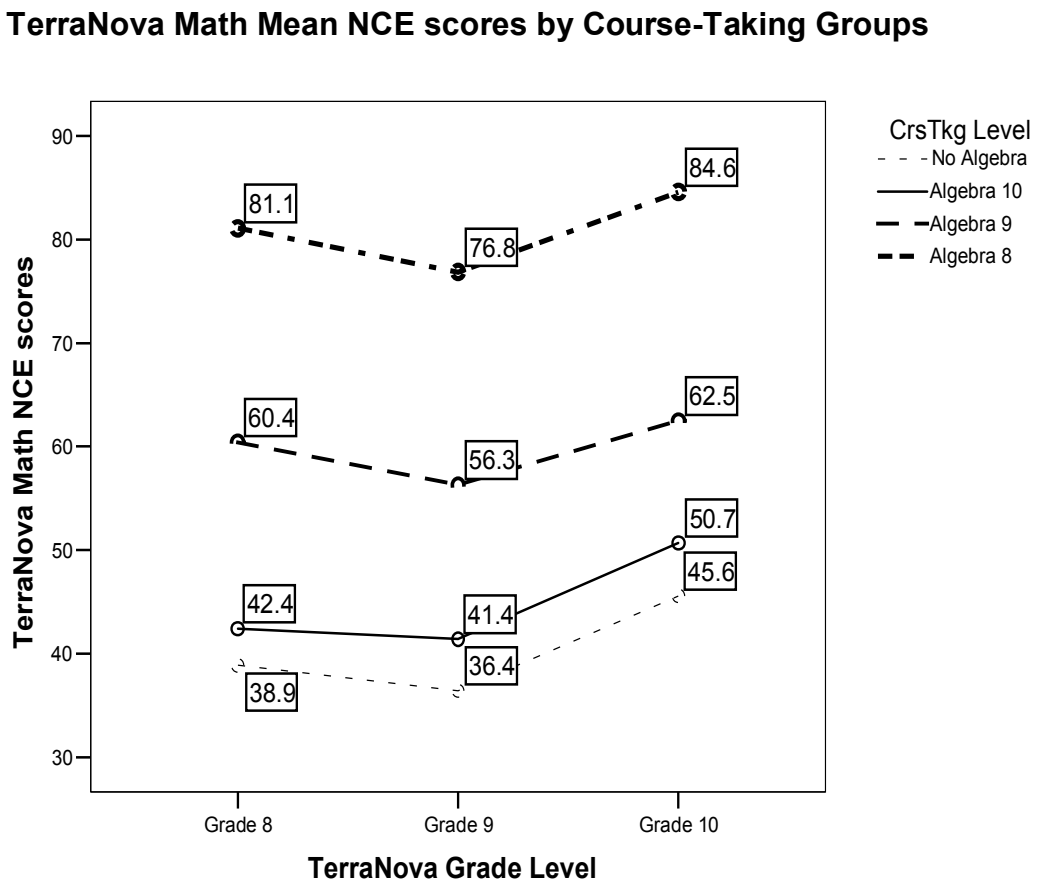

Figure 11. Mean Mathematics NCE scores for TerraNova mathematics tests for each course-taking group at grades 8, 9, and 10.

Results of the evaluation of the assumptions of normality of sampling distributions, linearity, and homogeneity of variance were satisfactory. The assumption of sphericity was met after the Huynh-Feldt correction ($\varepsilon = 0.90$) was applied. Using Wilks' λ, the results indicated a significant within-subjects *time* effect, Wilks' $\lambda = 0.81$, $F(2, 507) = 60.484$, $p < 0.001$. The *time by course-taking* interaction was not significant Wilks' $\lambda = 0.980$, $F(6, 1014) = 1.683$, $p = 0.122$. The between-subjects main effect of *course-taking* was significant, $F(3, 508) = 241.881$, $p < 0.001$. Figure 11 shows that the profiles for the course-taking levels are similar with students taking Algebra earlier

performing better than students who take Algebra later or not at all. In each course-taking group, the NCE scores dipped in grade 9 but increased beyond the grade 8 level scores by tenth grade. Although the plots are nearly parallel, the difference between performance by course-taking levels is not consistent across levels. Figure 11 shows that Levels 1 and 2 are approximately five points from one another at each grade level, Levels 2 and 3 have an average difference of approximately 15 points, while Levels 3 and 4 are approximately 21 points from one another at each grade level.

The results of the statistical analyses were reported in this chapter. Each of the seven hypotheses was accepted. Chapter V includes a summary of the study, discussion of the research findings, conclusions, implications of the findings, and recommendations for future research.

CHAPTER V

Summary of the Study

This study examined the relationship between two different measures of student achievement in mathematics and the course-taking behaviors of students, taking into account gender. Included in the student achievement data were individual student scores on the mathematics portion of the Missouri Assessment Program (MAP) tests in grades 8 and 10, and the TerraNova Multiple Assessments in grades 8 and 9. The construct of student achievement in mathematics is what both the MAP and the TerraNova mathematics tests purport to measure. The MAP is a criterion-referenced test with multiple choice (MC), constructed response (CR), and performance event (PE) items. The TerraNova Multiple Assessment is a norm-referenced test with both MC and CR items. Because the MAP test is intended to measure what students know and are able to do, the students are required to provide written responses to open-ended questions. Researchers have found a positive relationship between achievement levels in reading and mathematics (Abedi et al., 2001; Czujko & Bernstein, 1989). To examine student ability in reading and writing, student scores on the Communication Arts portion of the TerraNova in grades eight and nine were also included in the data analyses.

The primary purpose of the study was to examine the relationship between student achievement scores on the MAP mathematics tests in grades 8 and 10, TerraNova mathematics test in grade nine and mathematics course-taking behavior, especially the year of Algebra completion. The interaction effect of course taking and gender was also examined. The data used in this study were examined using regression and correlation techniques to describe relationships and to determine if factors could be identified that

would predict student performance on the MAP in grade 10. The subjects were students who were tested with each measure and enrolled in grade 10 in 1999-2000, 2000-2001, 2001-2002, or 2002-2003 in a suburban school system in Missouri.

The following research questions were addressed by this study.

1. Is there a relationship between gender and item type on the grade eight or grade 10 MAP mathematics tests? The MAP contains three item types: Multiple Choice (MC), Constructed Response (CR), and Performance Event (PE).
2. Is there a relationship between course taking, gender and content strand scores on the eighth grade MAP? (The MAP contains questions on six content strands: Number Sense, Geometry/Spatial Sense, Data Analysis/Probability, Mathematical Systems, and Discrete Mathematics)?
2. Is there a relationship between scores on the eighth grade or the tenth grade Missouri Assessment Program (MAP) test in mathematics and course-taking behavior in mathematics, taking into account gender?
4. Can the proficiency level(s) on the 10^{th} grade Missouri Assessment Program (MAP) be predicted by some combination of the factors of mathematics course taking, performance on eighth-grade MAP mathematics, TerraNova Comprehensive Tests of Basic Skills (CTBS) in Communication Arts, student grade point average (GPA) in mathematics for grades 8 through 10, race, or gender?
5. Is there a relationship between scores on the TerraNova mathematics test in grades 8, 9, and 10 and course-taking behavior in mathematics?

The literature review focused on three major themes relevant to this study. First was the fundamental theme of large-scale assessment as a public accountability measure. This theme was examined in several parts: the historical development of assessment in the United States since 1950; relationships among assessment, curriculum, and instruction; validity and reliability; item types (Multiple choice, Constructed response, Performance events); the use of assessments as accountability measures; and equity and bias. The MAP test is the accountability measure used in Missouri for both public school accreditation by the state and for the federal government accountability requirements of the No Child Left Behind initiative.

The second theme was course taking. This was examined in six parts including: opportunity to learn (OTL); the relationship of secondary school course taking to performance beyond high school; the mathematics pipeline; ability-grouping and tracking; the year of Algebra completion; and graduation requirements.

The third theme was gender, as it relates to performance in mathematics. Gender studies were examined in four parts that included the interaction of gender with the following: assessment, variability of scores, item types, and content strands. Studies that dealt with the interaction effects of two or more of these themes were also reviewed.

Findings

The first two hypotheses proposed a significant relationship between gender or item type, and performance on the MAP. The MAP tests include Multiple Choice, Constructed Response, and Performance Event items.

Hypothesis One was accepted. There was a statistically significant relationship between item type and performance. Regardless of gender, students scored the highest

percentage of points available on Multiple Choice items, followed by Constructed Response, then Performance Events. The males scored higher than the females on all three item types but the gender effect was not statistically significant in eighth grade. The profiles of the males and females were nearly parallel.

The analysis for Hypothesis Two led to similar conclusions in grade 10. Hypothesis Two was also accepted. Again the profiles were very similar with Multiple Choice scores being much higher than Constructed Response or Performance Event scores. However, the grade 10 analysis revealed a main effect of gender to be statistically significant, with males scoring significantly higher than females on the Multiple Choice items.

Hypothesis Three proposed a significant relationship between course-taking, gender, and student performance on items in the six content strands on the MAP tests in grade eight. Hypothesis Three was accepted. The analysis of student performance on the Geometry and Spatial Sense content strand was the only strand that yielded a statistically significant interaction effect of gender by course-taking. The main effect of gender was found to be statistically significant in the case of the Number Sense content strand and the Mathematical Systems content strand. The main effect of course taking was found to be statistically significant in the case of Number Sense, Data Analysis and Probability, Patterns and Relationships, Mathematical Systems, and Discrete Mathematics.

Hypothesis Four proposed a significant relationship between gender, course-taking behavior and performance on the grade eight MAP mathematics test. The students' grade eight TerraNova Language scores were used as a covariate in an attempt to focus on the effects of course taking and gender. This model explained 70% of the variance in

MAP scores. Hypothesis Four was accepted. Both gender and course-taking were statistically significant main effects. The interaction effect of gender by course taking was not statistically significant.

Hypothesis Five proposed a significant relationship between gender, course-taking behavior and performance on the grade ten MAP mathematics test. The student's grade nine TerraNova Language scores were used as a covariate in an attempt to focus on the effects of course taking and gender. This model explained 53% of the variance in MAP scores. Hypothesis Five was accepted. Both gender and course taking were statistically significant main effects. The interaction effect of gender by course taking was not statistically significant.

Hypothesis Six proposed that factors could be identified that would serve as significant predictors of MAP 10 performance. Hypothesis Six was accepted. An equation was developed using logistic regression. The significant predictor variables were MAP 8 mathematics T-scores, MATH GPA for grades 8 through 10, the factor of early Algebra taking (an Algebra course completed in grade eight), and gender. The success rate for predictions increased from 81.8% in the constant-only model to 91.4% in the model developed using these four predictors. Other factors that were included in the preliminary analysis were not found to be significant predictors of MAP 10 mathematics proficiency. Those factors were race, TerraNova Language scores in grade nine, and TerraNova Reading scores in grade nine.

Hypothesis Seven proposed a significant relationship between course taking and student performance on the norm-referenced TerraNova test over three consecutive years. The TerraNova is a portion of the MAP mathematics test that is developed by CTB

McGraw Hill. It includes both Multiple Choice and Constructed Response items. The TerraNova Multiple Assessments were given in grade nine. The math portion of this assessment also included Multiple Choice and Constructed Response items. The course-taking level group means for the TerraNova NCE scores from these three grade levels were analyzed for Hypothesis Seven. The most rigorous course-taking level was Algebra in grade eight, followed by Algebra in grade nine, followed by Algebra in grade 10 or no Algebra completed by the end of grade 10. The graphs of the four course-taking levels were nearly parallel but not separated by equal distances. The student group who took Algebra in grade eight had mean NCE scores an average of approximately 21 points higher than the student group who completed Algebra in grade nine (20.7, 20.5, 22.1). The Algebra in grade nine group scored an average of 13.5 points higher than the Algebra in grade 10 group, with the differences becoming smaller from grades eight through 10, (20.5, 14.9, 5.0). The two lowest groups, No Algebra and Algebra in grade 10 averaged only a 4.5 point difference between them (3.5, 5.0, 5.1). In each of the four groups, the NCE scores dropped in ninth grade and then in tenth grade rose above the level achieved in grade eight. The results of the analysis indicated a statistically significant main effect of course-taking.

Conclusions

Although specific test items were not available for analysis in this study, the organization of the reports of MAP results by item type and content strand allowed an analysis of performance on these factors by the independent variables of course-taking level and gender. Both males and females had their highest scores on the Multiple Choice portion of the MAP in grades eight and 10, followed by Constructed Response and

Performance Event items. This may be explained by the fact that Multiple Choice and some Constructed Response items are part of the TerraNova norm-referenced portion of the MAP. Either the same items or parallel items, which test the same content, are used year after year. Teachers know what to expect and can prepare students well for demonstrating mastery of this material. Performance Events, on the other hand, are not repeated annually. Teachers do not know from year-to-year what the questions will be or even which content strand will be tested by Performance Event items.

The analysis of performance on item types did not take into account different levels of ability in language and reading, so it is not clear whether Performance Event scores are related more to ability in Communication Arts or ability in Mathematics. Gender studies have shown that the gender gap in mathematics, favoring males, is narrowing; however, the gender gap, favoring females, in reading and writing persists (Coley, 2001; Fan & Chen, 1997; Gambell & Hunter, 1999; Han & Hoover, 1994; Hedges & Nowell, 1995; Hyde et al., 1990; Kleinfeld, 1998; Lee & Ware, 1986; Maccoby & Jacklin, 1974; McLure, 1998; Nowell & Hedges, 1998; Pallas & Alexander, 1983; Pomplun & Sundbye, 1999; Rebhorn & Miles, 1999; Ryan & Fan, 1996; Taylor et al., 1996; Wainer & Steinberg, 1992; Wilder & Powell, 1989; Willingham & Cole, 1997). The results of the item type analysis in this study did not support higher levels of performance by females on items involving a written response in mathematical settings. The results of the analyses at both grades 8 and 10 showed males outperforming females on all three item types, with statistically significant differences favoring males on Multiple Choice items in grade 10. Males' superior performance on Multiple Choice items has been reported in the research (McKendree, 2002; Myerberg, 1996). In fact,

many of the gender studies that form the body of research on gender differences in mathematics are focused on student achievement measures that include only Multiple Choice items.

The females in this sample outperformed males in TerraNova language at both grade 8 (Females: $\bar{x}$ = 62.13, Males: $\bar{x}$ = 57.53) and grade 9 (Females: $\bar{x}$ = 65.44, Males: $\bar{x}$ = 58.67). A closer inspection of TerraNova Language scores showed that females outperformed males at all course-taking levels, except Algebra in grade 8. Males scored slightly higher than females on Language in that course-taking group (grade 8 males: $\bar{x}$ = 74.77, grade 8 females: $\bar{x}$ = 74.28; grade 9 males: $\bar{x}$ = 79.13, grade 9 females: $\bar{x}$ = 76.52). Although females overall showed greater ability in language, this language ability was not correlated with higher scores on items that required a written response. One possible conclusion is that these items measure mathematics content more than language ability, which is consistent with the intent of the MAP mathematics test.

The content strand analysis for Hypothesis Three only looked at performance on the MAP eight mathematics test. Other researchers have suggested that differential course-taking outside of mathematics could contribute to performance on a test such as MAP and confound the analysis of results by math course-taking behavior (Metcalf, 2002). Although students in grade eight in this sample only had an opportunity to take either Algebra or Pre-Algebra, the Pre-Algebra group was split into the students who would take Algebra in grade nine and those who would not. The performance by the Algebra in grade 10 and No Algebra groups was adversely affecting the analysis of the performance by the total Pre-Algebra in grade eight group. Students in the two lowest course-taking levels are students who struggle in math, evidenced by an analysis of their

group mean mathematics GPA in grade eight. The mean GPA by course-taking levels in grade eight was significantly different for all groups, except groups one and two (Group 4 = 3.07, Group 3 = 2.59, Group 2 = 1.51, Group 1 = 1.16). The Hypothesis Three analyses for the six content strands showed that course taking mattered in every case. The results consistently showed student performance levels increasing with the level of course taking.

The TerraNova Language scores were not a part of this content strand analysis, so it is not possible to determine whether student performance on the content strands was a factor of course taking, language ability or both. Consistent with other research findings (Harris & Carlton, 1993; Lane et al., 1996), gender played a significant role in the Geometry and Spatial Sense content strand with males outperforming females. Some researchers (Willingham & Cole, 1997) have found that test items containing a figure favor males but there was no information available in the data for this study to indicate whether geometry items on the MAP contained a figure.

The Missouri Department of Elementary and Secondary Education (DESE) organized most of the Algebra content under the Mathematical Systems content strand. Although students with Algebra in grade eight would have more exposure to Algebra content, there is no evidence that the students who took Algebra in grade eight had any greater advantage on this Mathematical Systems content strand than on any other content strand. This may indicate that the district involved in this study is following the NCTM recommendation of teaching algebra concepts to all students. Also the MAP may not include items that can only be answered using formal algebraic methods.

The content strand analyses showed an interaction by gender at the lowest course-taking level for each content strand, except Mathematical Systems. For each of the other five content strands, males outperformed females except in the lowest course-taking level (the combination of No Algebra and Algebra in grade 10), where females outscored males by a small but significant margin. The data provide no clear explanation for this difference in performance by gender at the lower course-taking levels. It may be a function of the way the course-taking levels were defined for this study or a function of some treatment that was not a part of this study such as special education, tutoring, or such classroom interventions as collaborative teaching. The only content strands where gender was a significant main effect were Mathematical Systems and Number Sense. These two content strands had the highest overall mean scores of the six content strands, 64.1 and 64.6 respectively. Males scored significantly higher than females in both of these content strands. Gender played a role in the significant interaction effect of course taking by gender for Geometry and Spatial Sense. In all content strands, except Geometry, the difference between the mean male and mean female scores was approximately one point. In the case of Geometry the difference was almost five points. In addition, the overall scores for Geometry were the lowest with an overall mean of 47.8 and the range was the largest at 53.1 points.

Throughout this study, TerraNova Language was used as a proxy for ability for two reasons. IQ scores were not available for students in this sample and researchers have found a positive correlation between reading/language ability and mathematics performance (Abedi et al., 2001; Czujko & Bernstein, 1989). The analyses for Hypotheses Four and Five attempted to isolate the effects of course taking on the MAP

mathematics scores by using TerraNova Language as a covariate. Even with these language scores used as a covariate, the effects of course taking were statistically significant. In grade eight, this ANCOVA model which used gender and course-taking levels as independent variables explained over 70% of the variance in MAP 8 math scores, and in grade 10, the model explained 53% of the variance. Both gender and course taking were found to be statistically significant main effects at grades 8 and 10 even after using the TerraNova language scores as the covariate.

Many educational studies do not lead to results that can be generalized to other groups of students outside the sample in the investigation. This study is no exception. Hypothesis Six led to a finding of factors that can be used with this sample to significantly improve the prediction of performance on the MAP 10 mathematics test over a constant-only model. The specific results of this logistic regression may not be applicable to any other group outside this sample. However, the four factors found to be significant (MAP 8 performance, MATH GPA for grades 8 through 10, Algebra in grade 8, and gender) can give educators something upon which to focus in their efforts to help more grade 10 students score at the proficient level. Of these four factors, the one that seems most likely to be manipulated is the Algebra in grade eight. The factors of race, TerraNova language, and TerraNova reading were not found to be statistically significant predictors in the logistic regression.

The last analysis involved an examination of performance by course-taking group on the norm-referenced TerraNova mathematics test. The results revealed significant differences between all pairs of course-taking groups. An interesting finding was that the NCE scores dipped for each group in grade nine. There were no data collected for this

study that would explain this drop in NCE scores. Some possible explanations would be student adjustment to a high school setting in grade nine affecting the scores. Another possible explanation is that the TerraNova test in grade nine is a low-stakes test. In grades eight and 10, the TerraNova is a part of the MAP, which is a high-stakes test for teachers, schools, and districts because of its use as an accountability measure both at the state and federal level. TerraNova at the ninth grade level has four content areas in one test booklet. Scores were not disaggregated by math teacher in this district, and the test was given to ninth graders in the spring with the results coming back the next fall when students had moved on to new classes. There was no evidence that the TerraNova scores at the ninth-grade level were very useful to the students, teachers, or school district.

Another interpretation of these TerraNova results is that they are consistent with MAP results. Higher course-taking levels consistently and significantly led to higher scores on the TerraNova. Researchers have questioned the merit of using open-ended items because of the cost in time to administer and money to develop and score them (Behuniak & Tucker, 1992; Lukhele et al., 1994; Oescher et al., 1992; Pearson & Garavaglia, 2003; Visintainer, 2002). If a norm-referenced test can provide the same data as a criterion-referenced test, it may not be a responsible use of time or money to continue to administer a performance-based criterion-referenced test like the MAP. In the case of MAP, most of the students with the scores in the top two levels are those in the accelerated track. The same stratification exists for the TerraNova. These results indicate that either course taking or ability or both are leading to higher test scores. At the same time, both the state and federal accountability systems are demanding that schools increase the numbers of students with scores at the proficient levels. This point leads to

the discussion of the implications of these findings. “The fact that the same historically small group of students is still succeeding in the academic fast track does not diminish the need for major advances by other students (Clune, 1998, p. 149).”

Implications

This study joins many others that have found that course taking in mathematics is strongly and positively correlated with performance on mathematics assessments (Alexander & Pallas, 1984; Bohr, 1994; Jones et al., 1986; Sebring, 1985; Smith, 1996; Useem, 1990).

> There appears to be consensus among researchers that quantity of schooling is positively related to academic achievement. Whether achievement is measured by ACT, SAT, or tests developed for NELS and HSB, higher test scores are associated with spending more time in related course work. (Goertz, 1989, p. 7)

Educators have experimented with the concept of 'Algebra for all' with different results. Gamoran and Hannigan (2000) found that the benefit of taking high school Algebra is weaker for students with low test scores. The authors offered possible explanations for why low-scoring students might benefit less from taking Algebra. One is that they simply have less capacity to learn, another is that they are tracked into a less rigorous curriculum, and still another is that they are scheduled into regular Algebra classes where the instructional methods are not well suited to low achievers. Two possible methods offered for providing access to ‘Algebra for all’ were Equity 2000 (no longer an option) or what was referred to as a “stretch” curriculum that bridged the gap

between general mathematics and Algebra by using an integrated hands on approach (Gamoran & Hannigan, 2000).

The summary report for Equity 2000 (Harris, 1998) indicates that the participation rates in Algebra and Geometry increased at all of the sites. The passing rates for Algebra and Geometry may be interpreted as improving. Although the percent of students passing Algebra decreased at all the sites, since a greater number of students took the classes, the number passing increased in some cases. However, many students still failed despite increased efforts to provide support to struggling students.

Smith cautioned that policy changes to provide everyone with Algebra in grade eight would probably dilute the effects. Algebra eight might then be stratified to remedial Algebra, regular Algebra, or expert Algebra. She also stated that under the 'Algebra for all' in grade eight system, a course that would "credential" students would then be Algebra seven. She points out that the NCTM recommendation is that algebra *concepts* should be taught throughout grades 5-8. (NCTM, 1989, p. 102).

Missouri currently only requires two mathematics credits in the state graduation requirements. However, even with only two mathematics credits required by the state for graduation, Missouri has the highest percentage (89%) of students in the nation taking Algebra II or Integrated Mathematics 3 by graduation. This is an increase of 31% from 1990 (Blank & Langesen, 2003). While Missouri students are increasing the intensity and number of mathematics courses taken in grades 7 through 12, the average Missouri scores on the NAEP are not very different from the national average scores.

The MAP results reported at the state level are not disaggregated in a way that allows educators to analyze whether the students across the state with the higher course-

taking levels are getting most or all of the higher scores. There is evidence from the reports of NAEP results that as increasing numbers of students take higher-level courses, the aggregate scores for those high-level course-taking groups are lowered. However, that may still mean that a greater number of students earn higher scores. Other benefits to individual students of more rigorous course taking might not be measured by the MAP, NAEP, or the TerraNova. Students may earn higher ACT scores (www.act.org/news/releases/2003/8-20-03.html) or acquire valuable skills for use in the workplace or in the next level of schooling (Adelman, 1999; Long, 2003; Rose, 2001; Roth et al., 2001; Schiller & Muller, 2003).

An alternative implication to more rigorous course taking for all students, is a multi-level assessment system. In a presentation in 1997, Kilpatrick made the point that there is no reason to expect that standards-based assessments, regardless of their design can successfully accomplish their goals.

> When curricula have different goals, they can be compared either on the goals they share in common, in which case important things are not measured, or on the entire set of goals, in which case each curriculum is at a disadvantage on the goals it did not attempt… Legitimate comparisons can only be made on common goals, which necessarily fail to capture much of what makes each curriculum unique… If we want to know what mathematics our students are learning from the programs they are in, we need to use instruments that are sensitive to all facets of those programs. (Kilpatrick, 1997, pp. 5-6)

The findings from this study support research that points to a positive relationship between course-taking levels and student achievement in mathematics. The continued use of a single assessment for students with different course-taking behaviors as part of an accountability system raises questions that future researchers might explore.

Future Research

Many studies have examined relationships between student achievement and socioeconomic status (SES) or gender, factors that are outside school control. More studies should be done that examine the relationship between curriculum and student achievement, using gender and race as moderating variables. In examining student achievement in mathematics, researchers need to examine the effects of tracking and acceleration on student achievement. Since current mathematics assessments often require more reading and writing skills than was the case in the past, more studies should be conducted that examine the relationship between language ability, reading ability, and mathematics performance. It is difficult to disentangle ability or aptitude from experience with the curriculum. Studies that examine instructional practices related to achievement measured on a standards-based assessment may help teachers develop effective interventions for low achieving students.

Researchers in Missouri, and other states using performance-based assessments, should conduct studies that examine the relationship between these state-level accountability systems and external measures, such as ACT, SAT, or AP examinations.

On April 22, 2005, the Missouri Department of Elementary and Secondary Education announced that the graduation requirements for Missouri will be changed. Beginning with the class of 2010, Missouri will require three courses in mathematics,

starting with a course in Algebra. Missouri educators are currently participating in regional discussions to determine how to assess student achievement.

> The task force endorsed the idea of having all high school students take an exam, such as the ACT or SAT, with an "add-on component" to address Missouri's academic standards. All students would be required to take this exam in the eleventh grade, but there would not be a state-mandated passing score. (http://www.dese.mo.gov/news/2005/hstaskforce.htm)

Hopefully, these changes in assessment and graduation requirements will lead future researchers to examine the relationships between these more rigorous course-taking requirements and student performance on the new assessments. The state and the nation have data on ACT and SAT performance related to course-taking behavior, so the effect of the new Missouri graduation requirements on student achievement on ACT and SAT scores can easily be examined by future researchers. Since Missouri has just announced these changes and has yet to determine how Missouri assessments may change, it is too soon to say how these changes will fit into the national and state accountability for schools and districts.

References

Abedi, J., Lord, C., & Hofstetter, C. (2001). *Impact of selected background variables on students' NAEP math performance* (Working paper no. 2001-11). Washington, DC: National Center for Education Statistics.

Adelman, C. (1999). *Answers in the tool box: Academic intensity, attendance patterns, and bachelor's degree attainment.* (No. EDD00097). Washington, DC: U.S. Department of Education.

Alexander, K. L., & Pallas, A. M. (1984). Curriculum reform and school performance: An evaluation of the "New Basics." *American Journal of Education, 92*(4), 391-420.

Alexander, N. A. (2002). Race, poverty, and the student curriculum: Implications for standards policy. *American Educational Research Journal, 39*(3), 675-693.

Anderson, J. (2002). Gender-related differences on open and closed assessment tasks. *International Journal of Mathematical Education in Science and Technology, 33*(4), 495-503.

American Association of University Women. (1992). *How schools shortchange girls: A study of major findings on girls and education.* Washington, DC: AAUW Educational Foundation, The Wellesley College Center for Research on Women.

Applegate, K. D. (2003). *The relationship of attendance, socio-economic status, and mobility and the achievement of seventh graders.* (Doctoral dissertation, St. Louis University, 2003). *Dissertation Abstracts International, 64,* 2711.

Armstrong, J. M. (1981). Achievement and participation of women in mathematics: Results of two national surveys. *Journal for Research in Mathematics Education, 12*(5), 356-372.

Atanda, R. (2000). Do gatekeeper courses expand education options? *Education Statistics Quarterly, 1*(1). Retrieved August 21, 2004 from: http://nces.ed.gov/pubs99/1999303.pdf

Ayalon, H. (2002). Mathematics and science course taking among Arab students in Israel: A case of unexpected gender equality. *Educational Evaluation and Policy Analysis, 24*(1), 63-80.

Bartman, R. E. (1998). *Assessment standards for Missouri public schools*. Jefferson City MO: Missouri Department of Elementary and Secondary Education. Retrieved August 21, 2004 from: http://www.dese.state.mo.us/divimprove/assess/assessmentstandardsjune1998.doc

Behuniak, P., & Tucker, C. (1992). The potential of criterion-referenced tests with projected norms. *Applied Measurement in Education, 5*(4), 337-353.

Beller, M., & Gafni, N. (1996). The 1991 International Assessment of Educational Progress in mathematics and sciences: The gender differences perspective. *Journal of Educational Psychology, 88*(2), 365-377.

Benbow, C. P., & Stanley, J. C. (1980). Sex differences in mathematical ability: Fact or artifact? *Science, 210*, 1262-1264.

Berlak, H., Newmann, F. M., Adams, E., Archbald, D. A., Burgess, T., Raven, J., et al. (1992). *Toward a new science of educational testing and assessment*. Albany, NY: State University of New York Press.

Berry, L. (2003). Bridging the gap: A community college and area high schools collaborate to improve student success in college. *Community College Journal of Research and Practice, 27*(5), 393-407.

Bevan, R. (2001). Boys, girls and mathematics: Beginning to learn from the gender debate. *Mathematics in School, 30*(4), 2-6.

Bice, C. J. F. (2002). *The relationship between elementary parent involvement programs and secondary students' achievement and attendance.* (Doctoral dissertation, St. Louis University, 2002). *Dissertation Abstracts International, 63,* 1259.

Bielinski, J., & Davison, M. L. (1998). Gender differences by item difficulty: Interactions in multiple-choice mathematics items. *American Educational Research Journal, 35*(3), 455-476.

Bielinski, J., & Davison, M. L. (2001). A sex difference by item difficulty interaction in multiple-choice mathematics items administered to national probability samples. *Journal of Educational Measurement. 38*(1), 51-77.

Bishop, J. H. (1996). Signaling the competencies of high school students to employers. In L. B. Resnick & J. G. Wirt (Eds.), *Linking school and work: Roles for standards and assessment* (pp. 79-124). San Francisco, CA: Jossey-Bass, Inc.

Blank, R., & Langesen, D. (2003). *State Indicators of Science and Mathematics Education 2003*. Washington, DC: Council of Chief State School Officers (CCSSO). Retrieved August 21, 2004 from the CCSSO web site: http://www.ccsso.org/Projects/science_and_mathematics_education_indicators/1085.cfm

Bohr, L. (1994). Courses associated with freshmen learning. *Journal of the Freshman Year Experience, 6*, 69-90.

Bottoms, G., & Presson, A. (2000). *Finishing the job: Improving the achievement of vocational students.* Atlanta, GB: Southern Regional Education Board.

Bratberg, W. D. (2002). *Comparison of student achievement based upon participation in the enhancing Missouri's instructional networked teaching strategies project as measured by the Missouri Assessment Program.* (Doctoral dissertation, University of Missouri, Columbia, 2002). *Dissertation Abstracts International, 63,* 1795.

Brody, L. E., & Blackburn, C. C. (1996). Nurturing exceptional talent: SET as a legacy of SMPY. In C. P. Benbow & D. Lubinski (Eds.), *Intellectual talent* (pp. 246-265). Baltimore, MD: Johns Hopkins University Press.

Brosnan, M. J. (1998). The implications for academic attainment of perceived gender-appropriateness upon spatial task performance. *British Journal of Educational Psychology, 68*, 203-215.

Burkam, D. T., & Lee, V. E. (2003). *Mathematics, foreign language, and science coursetaking and the NELS:88 transcript data* (No. NCES 2003-01). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Burton, N. (1996). Have changes in the SAT affected women's mathematics performance? *Educational measurement: Issues and Practice, 15*(4), 5-9.

Carleton, D. (2002). *Students' guide to landmark congressional laws on education*. Westport, CT: Greenwood Press.

Catsambis, S., Mulkey, L. M., & Crain, R. L. (2001). For better or worse? A nationwide study of the social psychological effects of gender and ability grouping in mathematics. *Social Psychology of Education, 5*(1), 83-115.

Clune, W. H. (1998). The 'Standards Wars' in perspective. *Teachers College Record, 100*(1), 4.

Clune, W. H., & White, P. A. (1992). Education reform in the trenches: Increased academic course taking in high schools with lower achieving students in states with higher graduation requirements. *Educational Evaluation and Policy Analysis, 14*(1), 2-20.

Cohen, D. K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis, 12*(3), 311-329.

Cohen, J. (1969). *Statistical power for the behavioral sciences*. New York, NY: Academic Press.

Coley, R. J. (2001). *Differences in the gender gap: Comparisons across racial/ethnic groups in education and work*. Princeton, NJ: Educational Testing Service.

Cooper, B. (1998). Using Bernstein and Bourdieu to understand children's difficulties with "realistic" mathematics testing: An exploratory study. *International Journal of Qualitative Studies in Education, 11*(4), 511-532.

Cooper, B., & Dunne, M. (2000). *Assessing children's mathematical knowledge: Social class, sex and problem-solving*. Buckingham, England: Open University Press.

Cuban, L. (1984). *How teachers taught: Constancy and change in American classrooms 1890-1980*. New York, NY: Longman, Inc.

Czujko, R., & Bernstein, D. (1989). *Who takes science? A report on student coursework in high school science and mathematics*. New York, NY: American Institute of Physics.

Darling-Hammond, L. (1985). Beyond standardization: State standards and school improvement. *The Elementary School Journal, 85*(3), 315-336.

Davenport, E. C., Davison, M. L., Kuang, H., Ding, S., Kim, S.-K., & Kwak, N. (1998). High school mathematics course-taking by gender and ethnicity. *American Educational Research Journal, 35*(3), 497-514.

English, F. W. (2000). *Deciding what to teach and test: Developing, aligning, and auditing the curriculum*. Thousand Oaks, CA: Corwin Press, Inc.

Fairman, J. C. (1999). *Policy and practice: The tension between assessment reform in Maine and Maryland and teachers' practice in middle-school mathematics.*: *Dissertation Abstracts International, 60*, 0370.

Fan, X., & Chen, M. (1997). Gender differences in mathematics achievement: Findings from the National Education Longitudinal Study of 1988. *Journal of Experimental Education, 65*(3), 229-242.

Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research, 62*(1), 61-84.

Feingold, A. (1994). Gender differences in variability in intellectual abilities: A cross-cultural perspective. *Sex Roles, 30*(1-2), 81-92.

Finn, J. D., Gerber, S. B., & Wang, M. (2002). Course offerings, course requirements, and course taking in mathematics. *Journal of Curriculum and Supervision, 17*(4), 336-366.

Friedman, L. (1995). The space factor in mathematics: Gender differences. *Review of Educational Research, 65*(1), 22-50.

Fuchs, L. S., Fuchs, D., Karns, K., Hamlett, C. L., & Katzaroff, M. (1999). Mathematics performance assessment in the classroom: Effects on teacher planning and student problem solving. *American Educational Research Journal, 36*(3), 609-646.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction*. (6th edition). White Plains, NY: Longman Publishers.

Gambell, T. J., & Hunter, D. M. (1999). Rethinking gender differences in literacy. *Canadian Journal of Education, 24*(1), 1-16.

Gamoran, A. (1987). The stratification of high school learning opportunities. *Sociology of Education, 60*(3), 135-155.

Gamoran, A., & Hannigan, E. C. (2000). Algebra for everyone? Benefits of college-preparatory mathematics for students with diverse abilities in early secondary school. *Educational Evaluation and Policy Analysis, 22*(3), 241-254.

Goals 2000: Educate America Act of 1994, 20 U.S.C.S.§5801. 108 Stat. 125 (1994).

Goertz, M. E. (1989). *Course-taking patterns in the 1980s*. New Brunswick, NJ: Center for Policy Research in Education.

Gonzalez, E. J., O'Connor, K. M., & Miles, J. A. (2001). *How well do advanced placement students perform on the TIMSS advanced mathematics and physics tests?* Boston MA: Lynch School of Education. Boston College.

Gordon, S. P., & Reese, M. (1997). High-stakes testing: Worth the price? *Journal of School Leadership, 7*(4), 345-368.

Hall, N. R. (2001). *Effects of algebra I grade level on mathematics course selection, value, and self-confidence.* (Doctoral dissertation, George Mason University, Fairfax, Virginia, 2001). *Dissertation Abstracts International, 62,* 2996.

Halpern, D. F. (2000). *Sex differences in cognitive abilities* (3rd ed.). Mahwah, NJ: L. Erlbaum Associates.

Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed-response science test. *Applied Measurement in Education, 12*(4), 211-235.

Han, L., & Hoover, H. D. (1994, April 5-7, 1994). *Gender differences in achievement test scores.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Harris, A. M., & Carlton, S. T. (1993). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. *Applied Measurement in Education, 6*(2), 137-151.

Harris, C. D. (1998). *Equity 2000 and district change: Signs of success* (Final report. Program evaluation report No. FR-EADD-98-44). New York, NY: The College Board.

Harris, R. B., & Kerby, W. C. (1997). Statewide performance assessment as a complement to multiple-choice testing in high school economics. *The Journal of Economic Education, 28*, 122-134.

Haury, D. L., & Milbourne, L. A. (1999). *Should students be tracked in math or science?* (ERIC Digest, ED433217). Washington, DC: Office of Educational Research and

Improvement. Retrieved August 21, 2004 from: http://www.ericfacility.net/databases/ERIC_Digests/ed433217.html

Heavner, E. M. (2002). *The effects of full-day or half-day kindergarten attendance, gender, and socioeconomic status on third grade communication arts achievement.* (Doctoral dissertation, University of Missouri, St. Louis, 2002). *Dissertation Abstracts International, 63,* 3431.

Hedges, L. V., & Friedman, L. (1993a). Computing gender difference effects in tails of distributions: The consequences of differences in tail size, effect size, and variance ratio. *Review of Educational Research, 63*(1), 110-112.

Hedges, L. V., & Friedman, L. (1993b). Gender differences in variability in intellectual abilities: A reanalysis of Feingold's results. *Review of Educational Research, 63*(1), 94-105.

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269*(5220), 41-45.

Hill, H. C. (2001). Policy is not enough: Language and the interpretation of state standards. *American Educational Research Journal, 38*(2), 289-318.

Hoover, J. P. (1998). *The impact of the Pennsylvania system of school assessment on instructional practices*. (Doctoral dissertation, University of Pittsburgh, 1998). *Dissertation Abstracts International, 60,* 399.

Horn, L. (1990). *Trends in high school math and science course taking: Effects of gender and ethnicity.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Horn, L., & Bobbitt, L. (2000). *Mapping the road to college: First-generation students' math track, planning strategies, and context of support* (Statistical analysis report No. NCES 2000-153). Washington, DC: National Center for Education Statistics.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*(2), 139-155.

Improving America's Schools Act of 1994, 108 Stat. 3518 (1994).

Jones, L. V., Davenport, E. C., Bryson, A., Bekhuis, T., & Zwick, R. (1986). Mathematics and science test scores as related to courses taken in high school and other factors. *Journal of Educational Measurement, 23*(3), 197-208.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425-461.

Kilpatrick, J. (October 4, 1997). *Five lessons from the new math era.* Paper presented at the Reflecting on Sputnik: Linking the past, present, and future of educational reform conference, Washington DC. Retrieved August 21, 2004 from the National Academies web site: http://www.nas.edu/sputnik/kilpatin.htm

Kleinfeld, J. S. (1998). *The myth that schools shortchange girls: Social science in the service of deception*. Washington, DC: Women's Freedom Network. Retrieved August 21, 2004, from the University of Alaska Fairbanks web site: http:// www.uaf.edu/northern/schools/myth.html

Kulik, J. A., & Kulik, C.-L. (1984). Effects of accelerated instruction on students. *Review of Educational Research, 54*(3), 409-425.

Kulik, J. A., Kulik, C.-L., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21*(2), 435-447.

Lane, S., Wang, N., & Magone, M. (1996). Gender-related differential item functioning on a middle-school mathematics performance assessment. *Educational Measurement: Issues and Practice, 15*(4), 21-27, 31.

Laughman, S. A. (2000). *The effects of a parent education program on student achievement, parental involvement and attitude.* (Doctoral dissertation, University of Missouri, St.Louis, 2000). *Dissertation Abstracts International, 61,* 3114.

Lee, J. (1998). The impact of content-driven state education reform on instruction. *Research in Middle Level Education Quarterly, 21*(4), 15-29.

Lee, V. E., & Bryk, A. S. (1988). Curriculum tracking as mediating the social distribution of high school achievement. *Sociology of Education, 61*(2), 78-94.

Lee, V. E., & Ware, N. C. (1986, April 16-20, 1986). *When and why girls "leak" out of high school mathematics: A closer look.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

LeSage, J. B. (2001). *High student achievement on the Missouri Assessment Program: Fourth grade teachers' perceptions of instructional practice based upon selected NCTM standards.* (Doctoral dissertation, St. Louis University, 2002). *Dissertation Abstracts International, 62,* 1707.

Levine, E. J. (1998). *Using performance assessment as a tool for reform in an urban school district*. (Doctoral dissertation, Fordham, 1998). *Dissertation Abstracts International. 60,* 0607.

Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education*. Boston, MA: Houghton Mifflin.

Linn, M. C., & Peterson, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development, 56*, 1479-1498.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*(1), 1-16.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Long, W. J. (2003). *Mathematics placement and mathematics achievement in the community college.* (Doctoral dissertation, University of Missouri, St. Louis, 2003). *Dissertation Abstracts International, 64,* 1572.

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. (2000). A longitudinal assessment of antecedent course work in mathematics and subsequent mathematical attainment. *Journal of Educational Research, 94*(1), 16-28.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University Press.

Madaus, G. F. (1994). A technological and historical consideration of equity issues associated with proposals to change the nation's testing policy. *Harvard Educational Review, 64*(1), 76-95.

Mayer, D. P. (1998). Do new teaching standards undermine performance on old tests? *Educational Evaluation and Policy Analysis, 20*(2), 53-73.

McKendree, J. (2002). *Negative marking and gender bias*. Retrieved February 7, 2004, 2003, from http://caacentre.lboro.ac.uk/resources/faqs/negative_marking1.shtml

McKnight, C. C., Crosswhite, F. J., Dossey, J. A., Kifer, E., Swafford, J. O., Travers, K. J., et al. (1987). *The underachieving curriculum: Assessing U.S. school mathematics from an international perspective*. Champaign, Il: Stipes Publishing Company.

McLure, G. T. (1998). High school mathematics course taking and achievement among collegebound students: 1987-1996. *NASSP Bulletin, 82*(598), 110-118.

McLure, G. T., Boatwright, M., McClanahan, R., & McLure, J. W. (1998, April 13-17, 1998). *Trends in high school mathematics course taking and achievement by gender, race/ethnicity, and class 1987-1997.* Paper presented at the 1998 annual meeting of the American Educational Research Association, San Diego.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11*(1), 3-9.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement: Issues and Practice, 8*(1), 14-22.

Metcalf, L. A. (2002). *Curriculum-sensitive assessment: A psychometric study of tracking as a distributor of opportunity to learn high school mathematics.* (Doctoral dissertation, University of Illinois, Urbana-Champaign, 2002). *Dissertation Abstracts International, 63,* 3842.

Miller, L. D., & Mitchell, C. E. (1994). Evaluating achievement in mathematics: Exploring the gender biases of timed testing. *Education, 114*(3).

Missouri Department of Elementary and Secondary Education. (1996). *Missouri's framework for curriculum development in mathematics K-12*. Jefferson City, MO: MO DESE.

Missouri Department of Elementary and Secondary Education. (2001). *Supplement to the Curriculum Frameworks*. Jefferson City, MO: MO DESE. Available at the DESE web site: http://www.dese.state.mo.us/divimprove/curriculum/frameworks/supplement/math1.html

Moses, M. S., Howe, K. R., & Niesz, T. (1999). The pipeline and student perceptions of schooling: Good news and bad news. *Educational Policy, 13*(4), 573-591.

Moses, R. P., & Cobb, C. E. (2001). *Radical equations: Math literacy and civil rights*. Boston, MA: Beacon Press.

Moss, P. A., & Schutz, A. (2001). Educational standards, assessment, and the search for consensus. *American Educational Research Journal, 38*(1), 37-70.

Muthen, B., Huang, L.-C., Jo, B., Khoo, S.-T., Goff, G. N., Novak, J. R., et al. (1995). Opportunity-to-learn effects on achievement: Analytical aspects. *Educational Evaluation and Policy Analysis, 17*(3), 371-403.

Myerberg, N. J. (1996, April 8-12, 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.

National Center for Education Statistics. (2002a). *The condition of education*. Washington, DC: U.S. Dept. of Education, Office of Educational Research and Improvement.

National Center for Education Statistics. (2002b). *Digest of education statistics*. Washington, DC: U.S. Dept. of Education, Office of Educational Research and Improvement.

National Center for Education Statistics. (2003). *The condition of education.* Washington, DC: U.S. Dept. of Education, Office of Educational Research and Improvement.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: U.S. Department of Education.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston VA: author.

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics*. Reston VA: author.

National Council of Teachers of Mathematics. (1995). *Assessment standards for school mathematics*. Reston VA: author.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston VA: author.

National Research Council. (2002). *Investigating the influence of standards: A framework for research in mathematics, science, and technology education*. Washington, DC: National Academy Press.

No Child Left Behind Act of 2001, 20 U.S.C.§ 6301. 115 Stat. 1425 (2002).

Nowell, A., & Hedges, L. V. (1998). Trends in gender differences in academic achievement from 1960 to 1994: An analysis of differences in mean, variance, and extreme scores. *Sex Roles: A Journal of Research, 39*(1-2), 21-43.

Oakes, J., Ormseth, T., Bell, R., & Camp, P. (1990). *Multiplying inequalities: The effects of race, social class, and tracking on opportunities to learn mathematics and science*. Santa Monica, CA: Rand Corporation.

Oescher, J., Kirby, P. C., & Paradise, L. V. (1992). Validating state-mandated criterion-referenced achievement tests with norm-referenced test results for elementary and secondary students. *Journal of Experimental Education, 60*(2), 141-150.

O'Neil, H. F., Abedi, J., Lee, C., Miyoshi, J., & Mastergeorge, A. (2001). *Monetary incentives for low-stakes tests.* (Research and development report No. NCES 2001024). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Pallas, A. M., & Alexander, K. L. (1983). Sex differences in quantitative SAT performance: New evidence on the differential coursework hypothesis. *American Educational Research Journal, 20*(2), 165-182.

Partenheimer, P. R., & Miller, S. K. (2001, September, 2001). *Eighth grade Algebra acceleration: A case study of longitudinal effects through the high school pipeline.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, Washington.

Pearson, P. D., & Garavaglia, D. R. (2003). *NAEP validity studies: Improving the information value of performance items in large scale assessments* (No. NCES 2003-08). Washington, DC: U.S.Department of Education, National Center for Education Statistics.

Pelavin, S. H., & Kane, M. (1990). *Changing the odds: Factors increasing access to college*. New York, NY: College Entrance Examination Board.

Pomplun, M. & Sundbye, N. (1999). Gender differences in constructed response reading items. *Applied Measurement in Education, 12*(1), 95-109.

Popham, W. J. (2001). *The truth about testing: An educator's call to action.* Alexandria, VA: Association for Supervision and Curriculum Development.

Preece, P.F.W., Skinner, N.C., & Riall, R. A. H. (1999). The gender gap and discriminating power in the National Curriculum Key Stage three science assessments in England and Wales. *International Journal of Science Education, 21*(9), 979-987.

Rebhorn, L. S., & Miles, D. D. (1999). High-stakes testing: Barrier to gifted girls in mathematics and science? *School Science and Mathematics, 99*(6), 313-319.

Reckase, M. D. (1999). *An analysis of the assessment requirements mandated by IASA Title I legislation 1*. Retrieved August 21, 2004, from Michigan State University, East Lansing, Michigan web site: http://ed-web3.educ.msu.edu/reports/ed-researchrep/00/00may-report1.htm

Renzulli, J. (2004). Expanding the umbrella: An interview with Joseph Renzulli. *Roeper Review, 26*(2).

Resnick, L. B. (1987). *Education and learning to think.* Washington, DC: National Academy Press.

Rock, D. A., & Pollack, J. M. (1995a). *Mathematics course-taking and gains in mathematics achievement* (NCES Statistics in brief No. NCES 95-714). Washington, DC: National Center for Education Statistics.

Rock, D., & Pollack, J. (1995b). *Psychometric report for the NELS:88 base year (1988) through second follow-up (1992)*. Washington, DC: National Center for

Education Statistics. Retrieved August 21, 2004 from: http://nces.ed.gov/pubs95/95382.pdf

Rose, H. (2001). *Issues in education: Math curriculum and earnings, test score gaps, and affirmative action.* (Doctoral dissertation, University of California, San Diego, 2001). *Dissertation Abstracts International, 62,* 2200.

Roth, J., Crans, G. G., & Carter, R. L. (2001). Effect of high school course-taking and grades on passing a college placement test. *The High School Journal, 84*(2), 72-87.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing* (CSE Technical Report No. 566). Pittsburgh, PA: Achieve, Inc. Center for the study of evaluation (CSE).

Ryan, K. E., & Fan, M. (1996). Examining gender DIF on a multiple-choice test of mathematics: A confirmatory approach. *Educational Measurement: Issues and Practice, 15*(4), 15-20, 38.

Salmon, S. J. (1997). *Alternative math assessment: Teachers making sense of assessment in their classrooms. (*Doctoral dissertation, State University of New York, Buffalo, 1997). *Dissertation Abstracts International, 58,* 0115.

Schiller, K. S., & Muller, C. (2003). Raising the bar and equity? Effects of state high school graduation requirements and accountability policies on students' mathematics course taking. *Educational Evaluation and Policy Analysis, 25*(3), 299-318.

Schoenfeld, A. H. (1994). What do we know about mathematics curricula? *Journal of Mathematical Behavior, 13*(1), 55-80.

Sebring, P.A. (1985). *Course taking, achievement, and curricular policy.* (Doctoral dissertation, Northwestern University, 1985). *Dissertation Abstracts International, 46,* 2157.

Shakrani, S. (1996). *Eighth-grade Algebra course-taking and mathematics proficiency* (NAEPFACTS No. NCES 96-815). Washington, DC: National Center for Education Statistics.

Sherman, H. (1989). *A comparison of three methods of teaching rational number concepts to preservice teachers.* (Doctoral dissertation, University of Missouri, St. Louis, 1989). *Dissertation Abstracts International, 50,* 1205.

Smith, J. B. (1996). Does an extra year make any difference? The impact of early access to Algebra on long-term gains in mathematics attainment. *Educational Evaluation and Policy Analysis, 18*(2), 141-153.

Spade, J. Z., Columba, L., & Vanfossen, B. E. (1997). Tracking in mathematics and science: Courses and course-selection procedures. *Sociology of Education, 70*(2), 108-127.

Strong, S., & Sexton, L. C. (1996). Kentucky performance assessment of reading: Valid? *Contemporary Education, 67*(2), 102-106.

Taylor, P. J., Leder, G. C., Pollard, G. H., & Atkins, W. J. (1996). Gender differences in mathematics: Trends in performance. *Psychological Reports, 78*, 3-17.

Teitelbaum, P. (2003). The influence of high school graduation requirement policies in mathematics and science on student course-taking patterns and achievement. *Educational Evaluation and Policy Analysis, 25*(1), 31-57.

Thorndike-Christ, T. (1991). *Attitudes toward mathematics: Relationships to mathematics achievement, gender, mathematics course-taking plans, and career interests.* (ERIC Document Reproduction Service No. ED347066).

Tuma, J. E., & Gifford, A. (1990). *Higher graduation standards and their effect on the course-taking patterns of college- and non-college-bound high school graduates, 1969 to 1987*. (ERIC Document Reproduction Service No. ED318767).

U.S. Department of Education. (1997). *Mathematics equals opportunity* (White paper). Washington DC: Department of education.

U.S. Department of Education. National Center for Education Statistics. (1996). *Pursuing Excellence,* Washington, DC: U.S.Government Printing Office.

Useem, E. L. (1990, April 16-20, 1990). *Getting on the fast track in mathematics: School organization's influences on math track assignment.* Paper presented at the Annual Meeting of the American Educational Research Association, Boston.

Vinovskis, M. A. (1998). *Overseeing the Nation's Report Card: The creation and evolution of the National Assessment Governing Board (NAGB)*. Washington, DC: National Assessment Governing Board. Retrieved August 21, 2004 from: http://www.nagb.org/pubs/95222.pdf

Visintainer, C. (2002). *The relationship between two state-mandated, standardized tests using the norm-referenced TerraNova and the criteria-referenced, performance assessment developed for the Maryland School Performance Assessment Program (MSPAP).* (Doctoral dissertation, Wilmington College, Deleware, 2002). *Dissertation Abstracts International, 63,* 0915.

Wainer, H., & Steinberg, L. (1992). Sex differences in performance on the mathematics section of the Scholastic Aptitude Test: A bidirectional validity study. *Harvard Educational Review, 62*(3), 323-336.

Wang, J. (1999). Opportunity to learn, language proficiency, and immigrant status effects on mathematics achievement. *Journal of Educational Research, 93*(2), 101-111.

Ware, M., Richardson, L., & Kim, J. J. (2000). *What matters most in urban school reform. How reform works: An evaluative study of National Science Foundation's Urban Systemic Initiatives. Study Monograph No. 1.* Arlington VA: National Science Foundation. Directorate for Education and Human Resources.

Wentzel, K. R. (1988). Gender Diferences in Math and English Achievement: A Longitudinal Study. *Sex Roles, 18*(11/12), 691-699.

Wilder, G. Z., & Powell, K. (1989). *Sex differences in test performance: A survey of the literature* (No. 89-3). New York, NY: College Entrance Examination Board.

Wiley, D. E., & Yoon, B. (1995). Teacher reports on opportunity to learn: Analyses of the 1993 California Learning Assessment System (CLAS). *Educational Evaluation and Policy Analysis, 17*(3), 355-370.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahway, NJ: Lawrence Erlbaum Associates, Inc.

Wilson, L. D., & Zhang, L. (1999, April13-17). *A cognitive analysis of gender differences on constructed-response and multiple-choice assessments in mathematics.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Wise, L. L. (1985). Project TALENT: Mathematics course participation in the 1960s and its career consequences. In S. F. Chipman, L. R. Brush & D. M. Wilson (Eds.), *Women and mathematics: Balancing the equation* (pp. 25-58). Hillsdale, NJ: Lawrence Erlbaum.

Young, D. B. (1997). Science as inquiry: Transforming science education. In A. L. Costa & R. M. Liebman (Eds.), *Envisioning process as content: Toward a renaissance curriculum*. Thousand Oaks, CA: Corwin Press, Inc.

Appendix A

March 15, 2004

Superintendent
---------School District

Dear ----------,

I am writing to make a formal request to use ---------- School District student data in a study I will conduct as partial fulfillment of the requirements for the degree of Doctor of Education at the University of Missouri at St. Louis. The Institutional Review Board (IRB) at the university requires a letter of participation from the school district superintendent in order to approve my research study. The working title of my study is "Relationship between course-taking behavior, gender, and mathematics achievement on the Missouri Assessment Program."

My study fits into the "Continuous Improvement by Design," adopted by ---------- Board of Education on May 21, 2002 in several ways. I will provide valuable information to ---------- School District about how student participation in courses relates to student performance on the Missouri Assessment Program (see number 2 under "2002-2003 District Goals"). Under "Design Supports," this study relates to both "Curriculum Alignment" and "Data Analysis."

Although many factors are known to influence student achievement, some factors can be manipulated and others cannot. Such factors as gender, race, and socioeconomic status are fixed; however, curriculum and instruction are factors that can be manipulated and they are factors that educators constantly strive to optimize. The primary purpose of my study is to examine relationships between course-taking behavior and performance on both the Missouri Assessment Program (MAP) and the Terra Nova. I would also like to examine gender as a moderating variable. I would like to collect data on race; however, I may not be able to analyze the data using race as a factor because there may not be sufficient numbers of students from each racial subgroup in each of the course-taking levels.

The accountability of No Child Left Behind (NCLB) mandates that schools in the United States must do whatever it takes to help all children become proficient in Mathematics by 2014. Although Missouri reports disaggregated data at state and local levels, the data do not include information on course-taking. Generally students with greater interest and ability in mathematics choose more rigorous curriculum. In ----------, about 20-25% of eighth grade students take Algebra in eighth grade. Similar percentages exist across the state and nation. However, not all of these students are able to earn scores in the

Advanced range on the MAP. It is not known what the relationship is between student scores and coursework. It is also not known what interaction effects may exist between gender, race, and course-taking. Some of the questions I hope to explore are:

- Do all of the MAP 8 and MAP 10 Proficient and Advanced scores belong to students in the accelerated curriculum?
- Are there statistically significant relationships between MAP mathematics scores and Terra Nova language arts scores?
- Are there combinations of factors that allow a prediction of the MAP 10 mathematics scores? Are some of the contributing factors those that can be manipulated?
- Does early Algebra correlate to higher achievement on the MAP? If so, is that higher achievement evidenced equally in scores on each of the content strands tested?
- What are the relationships between performance on content strands, item-type, gender, and course-taking behavior?

---------- School Board Policy ------, part -- allows the use of student information in a study that has a purpose of improving instruction. This policy applies to my intended use for student data. I intend to collect individual student test score data from both the Missouri Assessment Program (MAP) and the Terra Nova, student grades and course data, student race and gender. Sources of this information will be Student Information System (SIS), student transcripts, student's permanent record files, district and building level reports of student test data, TestMate Clarity, and Clear Access. After student test scores have been matched to student transcript and demographic information, the data will be entered into a spreadsheet that can be used to import data into Statistical Package for the Social Sciences (SPSS) using student identification numbers rather than student names. Each student name will be matched with an Identification number assigned by me. Only my dissertation committee chair, Lloyd Richardson, Ph.D., and I will know the names of students that correspond to the identification numbers.

I anticipate the data collection process will be time consuming. I will only engage in this data collection activity outside the hours of my normal workday. When I spoke with you about my study last summer, you gave me verbal permission to use ---------school district student data. At the time, you requested that I discuss access to transcript data, permanent student records, and any other records archived at the high school with ----------, the high school principal. ---- gave me verbal permission at that time. I have secured and attached his written permission.

Neither individual students nor individual schools will be identified in the study. The district will be described as a suburban school district in Missouri. The description of the setting for the study will include describing socio-economic status of the district by placing the per pupil expenditure and percent of students receiving free and reduced lunch in the context of the state and the county. In other words, the district's per-pupil expenditure may be described as near the median for the state and in the lowest quartile for the county. Descriptions will be given in such a way as not to identify the district but to give the readers a context in which to understand the setting for the study. Socio-

economic status (SES) data on individual students will not be used but the SES of the district will be a proxy for the individual student SES data.

The ---------- School District Policy -------- states that "the information is destroyed when no longer needed for the purposes of the study." I will maintain the data in a secure place at my home for five years after the completion of the study. The data will only be used for this study. After five years, the data will be destroyed.

---------- School Board Policy ------, part -- requires that confidentiality be maintained. If you wish, I will share this ---------- policy with the members of the University of Missouri staff who will know that my study is being conducted using ---------- School District data. The only UMSL staff this would pertain to that I am aware of are the IRB staffmembers, who will have the copy of your letter of participation, and the members of my dissertation committee.

To offer more detailed information about my proposed study, I have attached a copy of the Institutional Review Board application and an abstract of my study. Of course, I will provide a copy of the completed study to you. I believe that my work will be beneficial to the district in providing a detailed analysis of data that can be used to optimize curriculum and instruction for all students. It is my hope that this information will contribute to increasing student achievement, which is not just something we all want to do but something we must do.

If you have any questions or concerns, please contact me. The best way to reach me is my cell phone number, which is ----------. I hope to obtain the required letter of participation very soon so that I can request IRB approval and the approval of my committee, and begin to collect the data.

Thank you for the support and assistance you and ---------- School District have given me as I continue learning!

Sincerely,

Geraldine Dressel Baumgart
Doctoral Candidate
University of Missouri St Louis

Atttachment: IRB request for exempt review
Research design proposal
---------- written permission to access data

Assigned Id Code: _____

Data Collection Instrument

Gender: M F **Race**: W B Other **Complete gr 10 in**: '00 '01 '02 '03 **Cohort_____** **Course taking level_____**

Year_____Mathematics Map 8: 1 2 3 4 5 Scale Score_____Raw Score:_____ TN NCE:_____

Points: MC____CR_____PE_____ %Strands: Number Sense____Geom/Spatial Sense_____Data _____

Patterns_____Mathematics'l systems_____Discrete_____

Year_____Mathematics Map 10: 1 2 3 4 5 Scale Score_____Raw Score:_____TN NCE:_____

Points: MC____CR_____PE_____ %Strands: Number Sense____Geom/Spatial Sense_____Data _____

Patterns_____Mathematics'l systems_____Discrete_____

TN 8 Reading Comp NCE: _____ **Reading Total**: _____

Lang Mech NCE:_____ Lang (Writing):_____ **Language Total NCE**:_____

Year_____TN 9 Reading Comp NCE: _____ **Reading Total**: _____

Lang Mech NCE:_____ Lang (Writing):_____ **Language Total NCE**:_____

Mathematics (Problem Solving) NCE:_____ Computation NCE:_____**Mathematics Total NCE**:_____

Mathematics Course taking and grades (All grades are unweighted) GPA (gr 8,9,10 mathematics)_____

	Pre-Algebra		*Algebra A*		*Algebra B*		*Algebra I*		*Geometry*		*Geometry Excel*		*Consumer Mathematics*		*Algebra II*		*Algebra II Excel*	
Grade 8																		
Grade 9																		
Grade 10																		

Appendix B

Table B1

Analysis of Variance for Item Type and Gender on grade 8 MAP mathematics test

Source	*df*	*MS*	*F*	*p*	η^2
		Between Subjects			
Gender	1	.265	2.008	.157	.004
Error	510	(.132)			
		Within Subjects			
Item Type	1.613	14.483	595.842	.000*	.539
Item x Gender	1.613	.039	1.625	.202	.003
Error	822.763	(.024)			

Note. Values enclosed in parentheses represent mean square errors .Item types are Constructed Response, Multiple Choice, and Performance Event.
The assumption of sphericity was met after the Huynh-Feldt correction was applied, epsilon = 0.807.
SPSS reports η^2 *as partial* η^2 *which is defined by* $SS_{effect}/(SS_{effect} + SS_{error})$.
$*p < 0.05$

Table B2

Analysis of Variance for Item Type and Gender on grade 10 MAP mathematics test

Source	*df*	*MS*	*F*	*p*	η^2
		Between Subjects			
Gender	1	.640	4.101	.043*	.008
Error	510	(.156)			
		Within Subjects			
Item Type	1.713	10.855	522.605	.000*	.506
Item x Gender	1.713	.002	.096	.881	.000
Error	873.621	(.021)			

Note. Values enclosed in parentheses represent mean square errors .Item types are Constructed Response, Multiple Choice, and Performance Event.
The assumption of sphericity was met after the Huynh-Feldt correction was applied, epsilon = 0.856.
SPSS reports η^2 *as partial* η^2 *which is defined by* $SS_{effecl}/(SS_{effect} + SS_{error})$.
*$p < 0.05$

Table B3

Analysis of Variance for gender, course-taking, and performance on the Number Sense content strand on the grade 8 MAP mathematics test

Source	*df*	*MS*	*F*	*p*	η^2
Gender	1	1003.537	4.770	.029*	.009
Course-taking	2	58669.133	278.868	.000*	.524
Gender * Course taking	2	574.158	2.729	.066	.011
Error	506	(210.383)			
Total	512				

R Squared = .533

Note. Values enclosed in parentheses represent mean square errors. Course-taking levels are Algebra completed in grade 8, Algebra completed in grade 9, and Algebra completed in grade 10 or not completed by the end of grade 10.

SPSS reports η^2 *as partial* η^2 *which is defined by* $SS_{effect}/(SS_{effect} + SS_{error})$.

*$p < 0.05$

Table B4

Analysis of Variance for gender, course-taking, and performance on the Geometry and Spatial Sense content strand on the grade 8 MAP mathematics test

Source	*df*	*MS*	*F*	*p*	η^2
Gender	1	4970.515	17.163	.000*	.033
Course-taking	2	70641.627	243.926	.000*	.491
Gender * Course taking	2	2500.146	8.633	.000*	.033
Error	506	(289.603)			
Total	512				

R Squared = .510

Note. Values enclosed in parentheses represent mean square errors Course-taking levels are Algebra completed in grade 8, Algebra completed in grade 9, and Algebra completed in grade 10 or not completed. SPSS reports η^2 *as partial* η^2 *which is defined by* $SS_{effect}/(SS_{effect} + SS_{error})$.

*$p < 0.05$

Table B5

Analysis of Variance for gender, course-taking, and performance on the Data Analysis and Probability content strand on the grade 8 MAP mathematics test

Source	*df*	*MS*	*F*	*p*	η^2
Gender	1	882.413	3.520	.061	.007
Course-taking	2	47806.784	190.714	.000*	.430
Gender * Course taking	2	568.026	2.266	.105	.009
Error	506	(250.673)			
Total	512				

R Squared = .439

Note. Values enclosed in parentheses represent mean square errors Course-taking levels are Algebra completed in grade 8, Algebra completed in grade 9, and Algebra completed in grade 10 or not completed. SPSS reports η^2 *as partial* η^2 *which is defined by* $SS_{effect}/(SS_{effect} + SS_{error})$.

$*p < 0.05$

Table B6

Analysis of Variance for gender, course-taking, and performance on the Patterns and Relationships content strand on the grade 8 MAP mathematics test

Source	*df*	*MS*	*F*	*p*	η^2
Gender	1	656.963	1.046	.307	.002
Course-taking	2	77679.466	123.649	.000*	.328
Gender * Course taking	2	1070.929	1.705	.183	.007
Error	506	(628.226)			
Total	512				

R Squared = .338

Note. Values enclosed in parentheses represent mean square errors Course-taking levels are Algebra completed in grade 8, Algebra completed in grade 9, and Algebra completed in grade 10 or not completed. SPSS reports η^2 *as partial* η^2 *which is defined by* $SS_{effect}/(SS_{effect} + SS_{error})$.

**p < 0.05*

Table B7

Analysis of Variance for gender, course-taking, and performance on the Mathematical Systems content strand on the grade 8 MAP mathematics test

Source	*df*	*MS*	*F*	*p*	η^2
Gender	1	3638.041	8.978	.003*	.017
Course-taking	2	49770.031	122.822	.000*	.327
Gender * Course taking	2	114.493	.283	.754	.001
Error	506	(405.219)			
Total	512				

R Squared = .330

Note. Values enclosed in parentheses represent mean square errors Course-taking levels are Algebra completed in grade 8, Algebra completed in grade 9, and Algebra completed in grade 10 or not completed. SPSS reports η^2 *as partial* η^2 *which is defined by* $SS_{effect}/(SS_{effect} + SS_{error})$.

**p < 0.05*

Table B8

Analysis of Variance for gender, course-taking, and performance on the Discrete Mathematics content strand on the grade 8 MAP mathematics test

Source	*df*	*MS*	*F*	*p*	η^2
Gender	1	32.206	.049	.824	.000
Course-taking	2	82052.979	126.002	.000*	.332
Gender * Course taking	2	966.647	1.484	.228	.006
Error	506	(651.204)			
Total	512				

R Squared = .342

Note. Values enclosed in parentheses represent mean square errors Course-taking levels are Algebra completed in grade 8, Algebra completed in grade 9, and Algebra completed in grade 10 or not completed. SPSS reports η^2 *as partial* η^2 *which is defined by* $SS_{effect}/(SS_{effect} + SS_{error})$.

$*p < 0.05$

Table B9

Analysis of Covariance results for gender, course-taking, grade 8 TerraNova Language performance (covariate) and performance on the MAP 8 mathematics test

Source	*df*	*MS*	*F*	*p*	η^2
		Between Subjects			
Gender	1	447.446	15.372	.000*	.040
Course-taking	3	1375.358	47.250	.000*	.277
Gender * Course taking	3	75.332	2.588	.053	.021
Covariate					
TerraNova Lang	1	3924.183	134.815	.000*	.267
Error	370	(29.108)			
Total	379				

R Squared = .704

Note. Values enclosed in parentheses represent mean square errors Course-taking levels are Algebra completed in grade 8, Algebra completed in grade 9, Algebra completed in grade 10, and Algebra not completed by the end of grade 10.

SPSS reports η^2 *as partial* η^2 *which is defined by* $SS_{effect}/(SS_{effect} + SS_{error})$.

$^*p < 0.05$

Table B10

Analysis of Covariance results for gender, course-taking, grade 9 TerraNova Language performance (covariate) and performance on the grade 10 MAP mathematics test

Source	*df*	*MS*	*F*	*p*	η^2
		Between Subjects			
Gender	1	1503.918	31.590	.000*	.059
Course-taking	3	2287.863	48.057	.000*	.223
Interaction	3	62.419	1.311	.270	.008
Covariate					
TerraNova Lang	1	2448.672	51.434	.000*	.093
Error	503	(47.608)			
Total	512				

R Squared = .532

Note. Values enclosed in parentheses represent mean square errors Course-taking levels are Algebra completed in grade 8, Algebra completed in grade 9, Algebra completed in grade 10, and Algebra not completed by the end of grade 10.

SPSS reports η^2 *as partial* η^2 *which is defined by* $SS_{effect}/(SS_{effect} + SS_{error})$.

*$p < 0.05$

Table B11

Logistic regression analysis of Proficiency on MAP 10 mathematics test predicted by T-scores on MAP 8 mathematics, Course-Taking, Math GPA for grades 8-10 and Gender

Variable	*B*	*SE*	*P*	*Exp(B)*
MAP 8 T-score	.153	.037	.000*	1.165
MATH GPA for Grades 8-10	1.238	.296	.000*	3.450
Course-Taking Level 4 (1=yes, 0=no)	1.513	.415	.000*	4.541
Gender (1=male, 0=female)	1.013	.349	.004*	2.753

Note: The MAP 8 T-scores variable is continuous. The MAP 8 scale scores were converted to Z-scores using means and standard deviations for the statewide population for the MAP test administrations included in this study (grade 8: 1998-2001).The MAP 8 Z- scores were then converted to T-scores. MATH GPA is the average of the available mathematics semester grades for each student in grades 8 through 10. If a student did not take math for one or more semesters, the GPA was computed using only the semesters where a math grade was given. Course-Taking Level 4 represents students who took Algebra in grade 8. The dependent variable, MAP 10 Proficiency, was dichotomous. One of the types of MAP scores reported is a MAP achievement level score(1 through 5). Missouri Department of Elementary and Secondary Education has identified Levels 4 and 5 as the desirable levels for all students. MAP 10 achievement level scores at levels 4 and 5 were considered Proficient and MAP 10 achievement level scores at levels 1, 2, 3 were considered Not Proficient.

Nagelkerke R square = 0.620
McFadden Pseudo R^2 = 0.504

Table B12

Summary of Repeated Measures *ANOVA Table for Grades 8 through 10 TerraNova mathematics NCE scores*

Source	*df*	*MS*	*F*	*p*	η^2
		Between Subjects			
Course Taking	3	102808.696	241.881	.000*	.588
Error	508	(425.038)			
		Within Subjects			
Time	1.796	6806.737	59.410	.000*	.105
Course x Time	5.388	239.454	2.090	.059	.012
Error	912.430	(114.573)			

Note. Values enclosed in parentheses represent mean square errors . Course-taking levels are Algebra completed in grade 8, Algebra completed in grade 9, Algebra completed in grade 10, and Algebra not completed by the end of grade 10. TerraNova NCE scores for grades 8 and 10 are from the TerraNova Survey portion of the MAP mathematics test. TerraNova NCE scores for grade 9 are from the mathematics portion of the TerraNova Multiple Assessments.

The assumption of sphericity was met after the Huynh-Feldt correction was applied, epsilon = 0.898

SPSS reports η^2 *as partial* η^2 *which is defined by* $SS_{effect}/(SS_{effect} + SS_{error})$.

**p < 0.05*