2-2013

Evaluating the Validity of Mcas Scores as an Indicator of Teacher Effectiveness

Jenna M. Copella
University of Massachusetts Amherst

EVALUATING THE VALIDITY OF MCAS SCORES AS AN INDICATOR OF TEACHER EFFECTIVENESS.

A Dissertation Presented

By

JENNA M. COPELLA

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

February 2013

School of Education
Educational Policy Research and Administration
Psychometric Methods, Educational Statistics, and Research Methods

EVALUATING THE VALIDITY OF MCAS SCORES AS AN INDICATOR OF TEACHER EFFECTIVENESS.

A Dissertation Presented

By

JENNA M. COPELLA

Approved as to style and content by:

____
Lisa A. Keller, Chair

____
Stephen G. Sireci, Member

____
George R. Milne, Member

____
Christine B. McCormick, Dean
School of Education

DEDICATION

For my children: Riley, Aidan, and Caitlin Copella. You bring meaning, purpose, and joy to my life. You are the three best things that have ever happened to me. I love you.

ACKNOWLEDGEMENTS

I would like to thank my advisor and committee chair, Dr. Lisa A. Keller. This work would not have been possible without her guidance and encouragement. Lisa, I would like to take this opportunity to say thank you. Thank you for sharing your time and knowledge with me, thank you for knowing when to push, thank you for understanding all of the things only another mother could understand, thank you for not judging, thank you for being my advisor, and thank you for being a friend. I hope you know how grateful I am and just how deeply I appreciate you.

I would also like to thank the rest of my department at UMASS; Ron Hambleton, Steve Sireci, Craig Wells, Jennifer Randall, April Zenisky, and Peg Louraine. They are a wonderful group of people. Steve Sireci introduced me to psychometrics, for which I will be forever grateful. I would never have found my way to UMASS without Steve. Ron Hambleton was a treasure trove of knowledge, anecdotes, and advice. I will hold dear many memories of Ron. Craig Wells, Jennifer Randall, and April Zenisky were a pleasure to work with and learn from every day. Peg Louraine does everything humanly possible to ease the way for the students in the program. The care and support these people provide their students is amazing. I am so lucky to have been a part of it. My life, both professionally and personally, has been enriched because of my time at REMP.

Thank you to Steve Sireci and George Milne for serving on my dissertation committee. Your comments, advice, and feedback have been invaluable.

Finally, thank you to my family. I have received so much love and support from everyone. I need to send a special thank you to my mother, Terri Sisson, who has supported me every step of this journey. I simply do not know what I would do without

her. And to my Aunty Jo, Joanne Coyle, who has always been there for me. My Mom and Aunty Jo have listened tirelessly to me talk about school and, more recently, my dissertation. They have both spent hours proofreading. And they both helped me with the dirtiest job of all – email addresses! Without them I would still be copying and pasting! They are incredible women. So to Mom and Aunty Jo… thank you for *everything*.

ABSTRACT

EVALUATING THE VALIDITY OF MCAS SCORES AS AN INDICATOR OF TEACHER EFFECTIVENESS

FEBRUARY 2013

JENNA M. COPELLA, A.A., SPRINGFIELD TECHNICAL COMMUNITY COLLEGE

B.A., WESTFIELD STATE UNIVERSITY

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Lisa A. Keller

The Massachusetts Department of Secondary and Elementary Education (DESE) has implemented an Educator Evaluation Framework that requires MCAS scores be used as a significant indicator of teacher effectiveness when available. This decision has implications for thousands of Massachusetts public school teachers. To date, DESE has not provided evidence to support the validity of using MCAS scores to make interpretations about teacher effectiveness. A review of the literature reveals much variation in the degree to which teachers use state-adopted content standards to plan instruction. The findings in the literature warrant investigation into teacher practice among Massachusetts public school teachers. The research questions for this study will be: 1.) Are there variations in the degree to which Massachusetts public school teachers use the Curriculum Frameworks to plan Math instruction?; and 2.) Is MCAS as an instrument sensitive enough to reflect variations in teacher practice in the student's scores? A survey of Massachusetts public school principals and Math teachers, grades three through eight, investigated the research questions. Survey results revealed that Massachusetts teachers use the Curriculum Frameworks to plan instruction to varying degrees. Survey results also suggest a lack of relationship between teacher practice

related to the use of the Curriculum Frameworks and student MCAS scores. These findings suggest MCAS scores may not be an appropriate indicator of teacher effectiveness; however, there are limitations to the study that require further investigation into these questions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In the summer of 2010, the Massachusetts Department of Elementary and Secondary Education (DESE) made public their intentions to use results of the Massachusetts Comprehensive Assessment System (MCAS) as a significant factor in the measure of teacher effectiveness (Vaznis, 2011). This decision, which enjoys support from the Massachusetts Teachers Association, marks a significant departure from the traditional practice of ignoring MCAS scores in the evaluation of teachers. The Massachusetts Commissioner of Elementary and Secondary Education, Mitchell Chester, stressed that the new system is not intended to be punitive; rather the intention is to aid schools in understanding how to effectively improve student performance, foster professional growth, and ensure that outstanding teachers are recognized and rewarded. MCAS scores have long been tied to accountability at the school level, as per the No Child Left Behind legislation of 2001. The consequences to schools for poor student performance are varied. Poorly performing public schools have seen loss of funding, loss of teachers and staff, and even risk being taken over by the state government as a result (Herman & Webb, 2007). Many of these assessments come with high stakes for the students as well. In many states, students may not be awarded a diploma if they fail to demonstrate some minimum required performance.

According to the Boston Globe (Vaznis, 2011) and a Massachusetts DESE task force report on teacher effectiveness (MA DESE, 2011a), a similar model of rewards and sanctions will be implemented for teachers. Based on a teacher’s perceived level of effectiveness (now significantly influenced by MCAS scores, recall), individual teachers may be rewarded financially, mandated to participate in professional development, awarded permanent status only if they are deemed effective within three years of hire, or they may be dismissed.

Massachusetts' decision to use MCAS scores as a measure of teacher effectiveness, after over 10 years of electing not to do so, likely stems from the desire to secure a portion of the funds available from Race to the Top. Race to the Top is a component of the American Recovery and Reinvestment Act implemented in 2009, legislation supported by Congress and President Obama. This legislation was an attempt to stimulate the American economy by infusing certain sectors with money meant to create jobs and improve practices in each sector and, by extension, benefit the community as a whole. Education was one of these sectors.

Race to the Top applications were judged on a point system in which points were awarded based on the degree to which each state's application addressed key areas of focus. The key areas were state success factors, standards and assessments, data systems to support instruction, great teachers and leaders, turning around the lowest achieving schools, and general selection criteria. Based on the point system, the great teachers and leaders section contained the largest proportion of points it was possible to earn for one category. One of the subcategories within this section was called "Improving teacher and principal effectiveness based on performance" and 42% of the total possible points for the section came from this subcategory alone. One requirement for earning points in this category/subcategory was that applicants incorporated student achievement data in the evaluation of teacher effectiveness. Massachusetts' application pledged to use student achievement data (e.g., MCAS results) as a considerable part of teachers' job reviews. Massachusetts' application was successful and the state was ultimately awarded 250 million dollars from the Race to the Top fund.

Unfortunately, the decision to use MCAS scores as a significant factor in the evaluation of teacher effectiveness is not without controversy. The literature on this topic is limited due to the fact that, prior to the Race to the Top initiative, assessment scores were not commonly used to evaluate teacher effectiveness. However, there are some examples in the literature in which researchers urge educators to proceed with caution before putting this model into practice (Phillips, 2009; Hinchey, 2010). A report from the Massachusetts Task Force of the Evaluation of Educators (March 2011a), which details the framework that includes measures of student achievement in the evaluation of teachers, offers no direct evidence from the literature supporting this practice. The task force report, in fact, repeatedly refers to the controversy surrounding this particular practice. This same report mentions Massachusetts' Race to the Top application several times, and was clearly influenced by the requirements within. One possible interpretation of this situation is that the use of MCAS data in the evaluation of teacher effectiveness is driven not by the desire on the part of educators to perpetuate good practice, but by the desire to secure government funds. It is troublesome to realize that the federal government has established a system within the realm of public education that not only endorses the use of questionable practices such as evaluating teacher effectiveness by using test scores, but actually forces individual states to participate in said questionable practices to receive grant funding.

There is no consensus in the field of education as to the best way to evaluate teacher effectiveness. There is some general agreement, though, that no one measure is representative of teacher effectiveness because teaching (and learning) is influenced by many factors (Hinchey, 2010). Research has shown many variables contribute to teacher effectiveness. The amount of education, experience, content knowledge, pedagogical knowledge, understanding of student development, classroom interaction, classroom activities, and teacher involvement in the school and community have all been linked to effectiveness when effectiveness is judged by student motivation, graduation rates, student behavior, and student well-being (Hinchey, 2010). Overemphasis on student achievement as measured by test scores ignores these other factors which, ironically, are the bricks in the path students walk to arrive at test scores.

There are methods that allow for evaluating teacher effectiveness based on the factors detailed above (Hinchey, 2010). Measuring those factors involves teacher observations, principal review, peer review, submission of portfolios, self-evaluation of classroom practices, student surveys, parent surveys, or some combination of these methods. These methods for evaluating teacher effectiveness also provide evidence of high quality teaching in the absence of high achievement scores, providing a more robust evaluation of teacher effectiveness.

The State of Massachusetts has developed an Educator Evaluation Framework (EEF) that incorporates several of the practices listed above (DESE, 2011a). The proposed EEF will focus on four standards for teachers: curriculum, instruction, and assessment; teaching all students; family and community engagement; and professional culture. The curriculum, instruction, and assessment standard, which is most directly related to the topic here, is defined as (DESE, 2011a, p. 16):

> The teacher promotes the learning and growth of all students through planning, instructional, and assessment activities that support a cycle of creating lessons focused on clear learning objectives, designing authentic and meaningful student assessments, analyzing student performance and growth, and continuously refining learning objectives.

The remaining standards address effective teacher practice as it relates to providing an environment conducive to learning based on high expectations, providing a safe and effective classroom, cultural proficiency, collaborative practice, and building effective partnerships with the community outside of the school.

There are three categories of evidence that will be considered in the evaluation of teacher effectiveness as defined by the above Standards. The categories are: a) multiple measures of student growth, learning, and achievement, b) judgments based on observation and artifacts of professional practice, and c) collection of additional evidence relevant to one or more Standard. There are multiple indicators within each of these categories that can be used to assess the degree to which a teacher is judged to be effective. The multiple measures of student growth, learning, and achievement category allows the use of measures of progress toward student learning targets, which are set between the teacher and an evaluator; MCAS growth measures compared to schools with comparable demographics; other statewide measures, such as the Massachusetts English Proficiency Assessment for English language learners; district-determined measures of student learning which can be compared across grades or subjects district-wide; and group measures that are aligned with the goals set by teams, across grade levels, departments, or schools.

Judgments based on observation and artifacts of professional practice must be based on either an observation system developed by the DESE or an observation system developed at the district level and approved by the DESE. The observation system must align with the EEF Standards, use the statewide rating scale approved by the Board of Elementary and Secondary Education, be informed by research and best practices, and capture important and discernible differences in teacher performance. The artifacts of professional practice can include lesson plans, unit plans, IEPs, and redacted written observation. Artifacts of professional practice can be thought of as teacher "products."

The collection of additional evidence relevant to one or more Standard will include evidence that the teacher compiles to demonstrate professional growth, contribution to the school and larger community, and the satisfaction of professional responsibilities. This could include evidence of professional development, student or teacher feedback, and evidence of peer collaboration.

The EEF described above is good news for Massachusetts teachers in that it incorporates many aspects of teacher evaluation endorsed by researchers in the field of education. Unfortunately, there is not a concrete definition of teacher effectiveness provided by the State of Massachusetts. Based on the task force report (DESE, 2011a), each Standard must be addressed, but exactly what evidence from each of these three categories will be required to demonstrate evidence of effective teaching will be determined at the district level by school committees (DESE, 2011b) with a significant contribution from "district and union leaders" (DESE, 2011a, p. 26). The exact EEF model used in each district will ultimately determine what constitutes an effective teacher by virtue of the evidence required to attain a teacher rating of exemplary, proficient, needs improvement, or unsatisfactory.

One requirement, though, for all districts will be the inclusion of multiple measures of student learning, growth, and achievement as a “significant factor” in all educator evaluations. This is where MCAS scores come into play. Again, there is no definition offered for what constitutes a significant factor and the weight of MCAS scores in determining an effective teacher may vary from district to district; however, all districts are mandated to give them significant weight. Multiple measures will take the form of a “combination of classroom, school, and district assessments and student growth percentiles where available” (DESE, 2011b). State regulations require that MCAS scores are considered as at least one of the two required measures of student learning, growth, and achievement (DESE, 2011b) when they are available.

MCAS scores will be used in the form of student growth percentiles (SGPs). SGPs are percentiles that are calculated based on a comparison of a student’s history of MCAS scores to other students with a similar history of MCAS scores (Chester, 2010). The use of SGPs was adopted by the State in an effort to use MCAS scores to separate growth from achievement (simply reported MCAS scores). Therefore, a student could demonstrate low achievement by scoring in the MCAS failing range, but still have high growth by advancing further in percentile rank as compared to other students with the same MCAS score history. The percentile essentially indicates a rate of change. The qualifier “where available” is necessary because only about 16% of teachers in Massachusetts teach at grade levels or in subject areas assessed by MCAS. However, when they are available, MCAS scores will factor prominently in the evaluation process in the form of student growth percentiles. And just so we do not lose sight of the actual number of teachers this regulation affects, please keep in mind that there were 68,754 public school teachers employed in the 2010 -2011school year (DESE, 2011c). 16% of that figure equals approximately 11,000 teachers to whom this regulation applies.

Given the support in the literature for factors such as those discussed above as good indicators of teacher effectiveness, why the sudden focus on student learning, growth, and achievement? Valid models for measuring those other factors are both time and cost intensive (Hinchey, 2010). However, all states currently have a standardized testing system in place in compliance with the No Child Left Behind legislation. It is really no surprise that Massachusetts elected to use MCAS scores. This course of action, which is being used in many other states as well, does not require the State to invest any additional time or resources in developing an indicator of teacher effectiveness. However, the MCAS is designed to be an end-of-year summative assessment of student achievement. In other words, a snapshot of a student's ability at one point in time. Any and all validity evidence accumulated to date has been in support of using MCAS scores for this purpose. Much of the literature surrounding this issue suggests that using achievement test scores to measure school or teacher effectiveness is not appropriate (Braun, 2004; Kupermintz, 2003; McCaffrey et al., 2003, 2004). Beyond the debate over the appropriateness of this practice lies a more important question: Can valid conclusions about teacher effectiveness be drawn based on inferences about MCAS scores? This is, at its heart, a validity issue.

The Standards for Educational and Psychological Testing (1999) (the Standards), the guiding source for good practice in the field of educational measurement, addresses the issue of validity. As the Standards tell us, "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 9). It seems prudent, then, to review the evidence provided in the literature to support the validity of inferences about teacher effectiveness based on student achievement (e.g., MCAS) scores.

As mentioned previously, all states have mandated standardized assessments. These assessments are aligned with state content standards. The content standards are usually developed by subject matter experts and, while the depth and breadth vary by state, represent what the students are expected to be taught and learn during the course of each academic year (Bhola, et al., 2003). This model of content standard to test alignment lends support to the validity of the interpretation of MCAS scores as an indicator of student achievement. The logic for using the scores as a measure of teacher effectiveness is that since we know what teachers are supposed to teach, we can measure the students to see if they have learned it.

This logic fails to address many issues known to influence student performance on assessments, such as socioeconomic status, characteristics of the student's home life, characteristics of the school district, and characteristics of the teacher (e.g., years of teaching experience) to name just a few. A trend emerges in the literature over the last decade or so in which researchers began to develop statistical models designed to account for influences on student performance and, more recently, to attempt to isolate the effect of a particular school, which can be extended to teachers. A brief review of the models currently in use follows.

Historically, the model used to evaluate growth or effectiveness at the school level has been the status model (Piche, 2007). This model compares examinee scores from one year to the next. The status model originated in the public health sector and is simply intended to track trends, such as the increase or decrease in infection rates of diseases (Piche, 2007). Researchers have recognized the shortcomings of using test scores to evaluate teachers without attempting to account for other outside influences on test scores. Status models may work for tracking student achievement, but they are not appropriate methods for evaluating effectiveness (Betebenner, 2009). The literature indicates that educators and psychometricians are attempting to control for the effects of outside influences by developing statistical models that are meant to, in part, isolate the contribution of the individual school or teacher to student achievement. These statistical models are referred to as growth models and value-added models. These models have enjoyed an increase in popularity and use within education over the last decade or so.

Growth models typically focus on student level characteristics, looking at whether the student is progressing and if the student appears to be on track to continue to make progress (O'Malley, McClarty, Magda, & Burling, 2011). These models focus on assessment scores over multiple years. Projection models, a variation of the growth model, typically label the student as on track or not (or some variation of such) and give a score that indicates expected gain. Growth models generally do not include student characteristics such as ethnicity or gender because expectations for growth should not be influenced by these factors. There are many types of growth models, but they are all similar in their efforts to measure student growth over time. In one example, Betebenner (2009), creator of the growth model currently being implemented in Massachusetts, transforms MCAS data, which is criterion-referenced, into normative data using his student growth percentile (2009). The current performance of a student is compared with other students that have a similar score history for the prior two years. Growth is then quantified in relation to the current performance of the other students. As mentioned previously, it is the SGPs that will be used as an indicator of teacher effectiveness in Massachusetts.

Value-added models (VAMs) are an attempt to delve even deeper into the meaning of a student's achievement/assessment scores by parceling the contribution of different factors, often using sophisticated statistical models (O'Malley, et al, 2011). VAMs look at the difference between predicted performance and observed performance, and attempt to link the different factors in the model to the student, the teacher, or the school (Soto, Sireci, Keller, & O'Malley, 2011). The focus of the VAMs, unlike the student-focused status or growth models, is the effect of outside factors on student achievement (O'Malley, et al, 2011; Soto, et al, 2011).

The question now is: How can we determine if valid inferences about teacher effectiveness can be made based on MCAS scores? There is a current movement to assess the validity of the models described above, especially VAMs (for a thorough review of this topic, please refer to Soto, et al, 2011); however, that work addresses a different question. The question at hand involves the validity of interpreting *student* achievement scores as an indicator of *teacher* effectiveness. One approach to answering this question is to construct a validation argument, as recommended by the Standards (1999). The Standards (1999) offer guidance in evaluating the validity of interpretations based on test scores by specifying the need to explicitly state the new proposed interpretation of test scores and provide a rationale to explain how the interpretation is relevant to the proposed use of the scores. According to the Standards (1999), the proposed interpretation should be used to guide test development, thus ensuring the conceptual framework of the test aligns with interpretations made based on the test scores. Unfortunately, the reality of the present situation does not allow for the thorough and complex treatment of validity called for in the Standards; however, some of the principles recommended to construct a validity argument can be adapted to aid in evaluating the validity of using MCAS scores as indicators of teacher effectiveness.

One aspect of constructing a validity argument that is recommended in the Standards (1999) and that would be useful for this study is the identification of propositions that support the proposed interpretations of test scores. The propositions are then evaluated based on empirical evidence, relevant literature, rational analysis, or some combination thereof. Some propositions that could be offered for using MCAS scores as indicators of teacher effectiveness are:

1. MCAS scores can tell us something about the teacher of the student taking the test.
2. MCAS scores are a valid indicator of student achievement.
3. Student growth percentiles provide information about the effectiveness of the student‘s current teacher.

4. MCAS is aligned with the Massachusetts Curriculum Frameworks (the Frameworks).
5. Massachusetts public school teachers are teaching the Frameworks (thus, the standards-based reform model described previously is operating properly in the context of MCAS).
6. MCAS scores are sensitive enough to reflect variations in teacher practice related to the use of the Frameworks to plan instruction.

And it is at this point we encounter two propositions that may be false, at least in some instances. Proposition 5, Massachusetts public school teachers are teaching the Frameworks, and proposition 6, MCAS scores are sensitive enough to reflect variations in teacher practice related to the use of the Frameworks to plan instruction, bear further investigation. There is evidence in the literature that may contradict the assumption that teachers are teaching the Frameworks. Furthermore, the history of MCAS pass rates and score distributions may raise questions regarding the sensitivity of MCAS scores to variations in teacher practice assumed in proposition 6.

The literature contains many studies that have shown that since the advent of standards-based education reform teachers have deviated from the content standards implemented at the State, or even Federal, level (American Federation of Teachers, 2009; Hamilton, Stecher, & Yuan, 2009; Sherin & Drake, 2010). Reasons cited for this practice range from feeling the content standards contain too much breadth and not enough depth of knowledge to the failure of teacher preparation programs to train teachers how to properly use content standards. If this deviation from the content standards is common practice amongst Massachusetts teachers, thus contradicting the assumptions implicit in proposition 5, what are the implications for using MCAS scores as a significant indicator of teacher effectiveness?

Given that the State defines an effective teacher as one who “promotes learning… focused on clear learning objectives…” (DESE, 2011a, p. 16), we then have to assume proposition 6, MCAS scores are sensitive enough to reflect variations in teacher practice related to the use of the Frameworks to plan instruction, is true. In practice, we should see a continuum of MCAS scores with high MCAS scores from students whose teachers do not deviate from using the Frameworks to plan instruction, lower MCAS scores from students whose teachers use the Frameworks to plan instruction to a lesser degree, and low MCAS scores from students whose teachers do not use the Frameworks to plan instruction.

Table 1 contains the MCAS pass rates and percentages of students scoring in each passing score category in English Language Arts (ELA) and Mathematics for the past five years. As you can see, the pass rates are high, particularly for ELA, and there is little variation in pass rates or classification rates across the five years.

Table 1. ELA and Math Pass/Classification Rates.

English Language Arts				
Year	Pass Rate (%)	Needs Improvement (%)	Proficient (%)	Advanced (%)
2011	92	23	52	17
2010	92	24	52	17
2009	92	25	51	16
2008	91	27	50	14
Mathematics				
Year	Pass Rate (%)	Needs Improvement (%)	Proficient (%)	Advanced (%)
2011	85	27	34	24
2010	85	27	33	26
2009	84	28	32	23
2008	83	28	31	24

Table created based on information from the Massachusetts DESE
http://profiles.doe.mass.edu/profiles/general.aspx?topNavId=1&orgcode=00000000&orgtypecode=0&

The score distribution is not a normal, bell-shaped distribution as you might expect given the large number of students taking MCAS exams each year. The score distribution is heavily, negatively skewed, with only 8% - 9% of ELA students and 15% - 17% of Math students receiving failing scores. This is a perfectly acceptable result given that MCAS tests are criterion-referenced tests; therefore, students are demonstrating their knowledge of a construct. However, if Massachusetts teachers deviate from using the content standards to the same degree as teachers across the country, the high pass rates and the classification of 55% - 69% of students into the two highest performance categories are suspect.

In addition to the high pass rates, the stability of the percent of students scoring in each performance category may be cause for concern. If MCAS scores truly reflect the degree to which teachers use the Frameworks to plan instruction, thus their effectiveness, these numbers suggest that there has been little to no variation in teacher practice related to the use of the Curriculum Frameworks to plan instruction over the last five years. Every year, though, experienced teachers have retired, newly licensed teachers have been hired, teachers have moved from one grade and/or subject to another, and changes have occurred at the administration level in many schools. Although it is possible that there have been no variations in teacher practice over the past five years related to the use of the Curriculum Frameworks to plan instruction, it seems unlikely that there have not been changes in teacher practice given the changing teacher population. The stability of performance category classifications are another indication that the sensitivity of MCAS scores to variations in teacher practice must be investigated. In order to use MCAS grades as indicators of teacher effectiveness, a link between MCAS scores and the degree to which teachers use the content standards to plan instruction must be established. The State has not provided any proof of this link; therefore, there is no evidence that the assumptions underlying proposition 6 are true. If MCAS scores are not sensitive to variations in teacher practice, they will not be an appropriate indicator of teacher effectiveness.

The purpose of this research is to examine the assumptions implicit in proposition 5 listed above, that Massachusetts teachers are using the Massachusetts Curriculum Frameworks (the Frameworks), the state-approved content standards, to plan instruction, thus teaching the content of the Frameworks and to examine the assumptions implicit in proposition 6, MCAS scores are sensitive to variations in teacher practice related to the use of the Frameworks to plan instruction. Through the use of surveys given to Massachusetts public school teachers, this research will attempt to ascertain the degree to which Massachusetts teachers adhere to or deviate from the content standards while planning classroom instruction, the appropriateness of MCAS as a measure of this, and the sensitivity of MCAS to variations in teacher practice, if variations exist. The results will inform a validity argument evaluating the interpretation of MCAS scores as an indicator of teacher effectiveness.

CHAPTER 2

LITERATURE REVIEW

A review of the most recent MCAS Technical Manual (2010) on the DESE website indicates that the State uses MCAS scores for the following purposes (Massachusetts DESE, 2010):

- measure student, school, and district performance in meeting the State's learning standards as detailed in the Massachusetts Curriculum Frameworks;
- improve student achievement and classroom instruction by providing diagnostic feedback regarding the acquisition of skills and knowledge;
- help determine English language arts (ELA), mathematics, and science and technology/engineering (STE) competency for the awarding of high school diplomas;
- hold schools and districts accountable for the yearly progress they make toward meeting the goal, set by the federal No Child Left Behind (NCLB) Act, that all students will become proficient in reading and mathematics (p. 1).

It is interesting to note that teacher effectiveness is not on the list. A further review of the technical manual reveals that the only validity evidence reported by the State pertains to the use of MCAS scores as an indicator of student achievement. The State offers evidence to support content validity, criterion-related validity, and consequential validity. This type of validity evidence supports the use of MCAS scores for evaluating competency for the awarding of diplomas.

Unfortunately, as MCAS tests are criterion-referenced snapshots of a student's knowledge of a construct at one moment in time, the validity evidence provided in the technical manual does not support the other claims in the Technical Manual (2010) or the new claim from the State that MCAS scores are indicators of teacher effectiveness. There is a common thread amongst the claims that MCAS scores can be used to evaluate school and district performance in meeting the State's learning standards, MCAS scores are evidence of school and district accountability towards the goal that all students will become proficient in reading and mathematics, MCAS scores can improve student achievement and classroom instruction by providing diagnostic feedback, and MCAS scores can be used as indicators of teacher effectiveness. The theme is using MCAS scores in the aggregate to evaluate persons (sometimes whole school districts of people) that did not take the test. Certainly the evidence provided in the Technical Manual (2010) to support content validity, criterion-related validity, and consequential validity does not address these proposed interpretations of MCAS scores.

On a more positive note, the 2010 Technical Manual devotes more attention to addressing validity evidence than previous years. This manual connects the alignment of the test blueprint and the Curriculum Frameworks, describes their procedures for ensuring adequate content coverage, and describes bias and DIF studies as indicators of evidence of content validity. The term “internal structure” encompasses reliability, item statistics, dimensionality studies, scaling and equating procedures. The internal structure is offered as further evidence contributing to the validity of the score interpretations. Criterion-related validity evidence is presented in the form of comparisons of MCAS results with other large-scale standardized assessments administered to some grade levels. This brief description of the 2010 Technical Manual is intended only to communicate an overview of the manner in which validity evidence is addressed; the manual provided a more comprehensive treatment of validity evidence. The treatment of validity in the 2010 Technical Manual is a commendable improvement over what was offered in previous years.

Using the 2010 MCAS Technical Manual as a guide, it seems the only claims about validity that can be addressed relate to student performance. A review of Technical Manuals for 2005, 2006, 2007, 2008, and 2009 is consistent with this finding. In each of these years the Technical Manual states "Evidence is presented in detail throughout this document to support inferences of student achievement of the learning standards of the *Massachusetts Curriculum Frameworks*, as measured by MCAS…" (MCAS Technical Manual, 2009, p. 144). Prior to 2005, there was very little attention devoted to validity in the Technical Manuals.

An additional document on the DESE website, a 2008 technical report called "Ensuring technical quality: Policies and procedures guiding the development of the MCAS tests", was reviewed to ascertain if there was information provided to suggest that MCAS scores are valid indicators of any of the above-stated uses. The technical report, as with the Technical Manuals, only addressed the validity of test scores used to evaluate student achievement. The following statement, taken from "Ensuring technical quality: Policies and procedures guiding the development of the MCAS tests" (2008), nicely summarizes the validity evidence provided by the State:

> The process is designed to produce MCAS tests that are aligned with the *Curriculum Frameworks* and that support valid inferences about student performance on the learning standards contained in the frameworks. (p. 9)

The manual goes on to describe the procedures used to create each form of the MCAS related to item development, alignment, and similar concerns.

Finally, there is a link on the DESE website labeled "Validity Studies" that directs the reader to the UMASS Center for Educational Assessment website. There are twenty-seven reports available providing details of various validity studies that have been conducted on the MCAS since 1998. The reports are not all validity studies, per se, but they are all investigations into topics that contribute to the overall validity of the interpretation of MCAS scores for measuring student achievement. The reports cover a wide range of topics including differential item functioning, consistency of test content, comparability of test forms, alignment studies, model fit, and equating procedures.

The purpose of reviewing the technical information and validity studies available about MCAS was to determine if the State of Massachusetts has provided any evidence to support that valid inferences about teacher effectiveness could be made based on MCAS scores. The answer is no, it has not. Therefore, this research will proceed using the guidelines found in The Standards (1999) to evaluate the validity of using MCAS scores as indicators of teacher effectiveness.

The Standards

The Standards (1999) are considered within the field of education to be the authority for good practice in educational testing. The Standards tell us that a validity argument must be constructed to support each proposed interpretation of test scores. The conceptual framework of the validity argument dictates the most appropriate evidence to support said argument. This means that there is not a prescribed set of steps one must follow to construct a validity argument for each proposed interpretation of test scores, but rather validity evidence may take the form of empirical evidence and professional judgment (The Standards, 1999).

The Standards cite five distinct types of evidence to support a validity argument; however, it should be noted that these are not types of *validity*. Validity is a unitary concept (The Standards, 1999, p. 11) and all evidence must be taken as a whole to evaluate the utility of a validity argument. The five sources of validity evidence fall into the following categories: evidence based on test content; evidence based on response processes; evidence based on internal structure; evidence based on relations to other variables; and evidence based on consequences of testing. A brief description of each source of validity evidence follows.

Evidence based on test content involves an evaluation of the relationship between the content included in a test and the underlying construct the test is intended to measure. This refers to all aspects of the test including administration, the content domain, the questions, and scoring. Alignment studies are a common source of validity evidence addressing test content (please see the Alignment Studies section of this chapter for further information). This category of validity evidence is, at the core, concerned with ensuring that the test actually measures what it is intended to measure. The Standards (1999) very specifically state that when using a test “for purposes other than that for which it was first developed, it is especially important to evaluate the appropriateness of the original content domain for the proposed new use (p. 12).

The previous section explained that there is evidence related to test content that supports the validity of interpretations of MCAS scores related to student achievement: the MCAS is aligned with the Frameworks. The new interpretations of MCAS scores as indicators of teacher effectiveness effectively change the content domain from what students know to what teachers teach. In light of this interpretation, there is no longer evidence that the test is actually measuring that which it is intended to measure because what teachers teach is filtered through the student. There is no way to separate out a positive or negative contribution from the student. That is, a teacher may not have taught the Frameworks but a good student may achieve a high score, or vice versa. There is no mechanism built in to the current evaluation system, including SGPs, that allows the parceling out of "credit" for the score. This is not a problem in the original use of MCAS scores where all of the "credit" is given to the student; however, the original validity evidence related to test content is not appropriate to support interpretations that MCAS scores are indicators of what teachers are teaching.

Evidence based on response processes involves making sure that the tasks or questions on a test actually require the examinee to engage in the processes of interest to provide correct responses, rather than arriving at a correct response some other way. For example, if mathematical reasoning is the construct of interest, the test questions must require examinees to demonstrate reasoning abilities rather than simply employing a previously memorized algorithm to solve an equation. This source of validity evidence is intended to ensure that the interpretations made about test scores are related to the construct of interest and not confounded by construct irrelevant variance.

Validity evidence based on response processes are related directly to the examinee taking the test. As the "examinees" in the new proposed interpretation of MCAS scores are the teachers, and the teachers are not taking the test, this category of validity evidence cannot be evaluated. As with validity evidence related to test content, the instrument that is intended to be used to provide information about teachers (MCAS) is filtered through a student. There is no way to evaluate the thought process of the teacher in this scenario. A completely different test, one in which the teachers actually answer the questions or that evaluates the teacher as s/he prepares lessons (e.g., a think aloud), would be needed to evaluate validity evidence based on response processes; therefore, this category of validity evidence cannot be addressed in this study.

Validity evidence based on internal structure evaluates the relationship between different components of a test. The components may be different subtests included in a test battery or the questions included on a test intended to be unidimensional. Reliability estimates and studies of differential item functioning are examples of methods used to evaluate the internal structure of a test.

As with validity evidence related to test content, DESE has provided validity evidence based on internal structure that supports the use of MCAS scores as indicators of student achievement. Also like validity evidence related to test content, this evidence does not support the use of MCAS scores as indicators of teacher effectiveness. The new interpretation of MCAS scores has effectively changed the content domain from what students know to what teachers teach; however, there is no change in the questions or the examinees, which means the wrong content domain is being assessed. Therefore, the validity evidence related to internal structure may indicate the MCAS is a reliable measure, but that does not support the argument that MCAS scores are valid indicators of teacher effectiveness because reliability is a necessary, *but not sufficient*, component of validity.

Validity evidence based on relations to other variables refers to what has, in the past, been termed discriminant and convergent validity. These terms are no longer popularly used, and are referred to in The Standards as discriminant and convergent evidence. Regardless of the name, this evidence refers to the relationship between the test of interest and other tests that measure either different constructs (divergent) or the same construct (convergent). That is to say, a test should not be highly related to a test purported to measure a different construct and it should be highly related to a test purported to measure the same construct. These relationships are often, though not always, evaluated through correlations. The purpose of evaluating this type of validity evidence is to evaluate how well the test predicts performance on some criterion. This directly impacts the interpretations made about test scores and on validity generalization. Validity generalization is the extension of validity evidence from one proposed use of a test to another. Validity evidence must be evaluated for each proposed use of a test to determine if it can be generalized to another context (The Standards, 1999).

It seems it would be possible to collect convergent evidence of the validity of MCAS scores. There are multiple methods currently used for teacher evaluation. A study could be constructed to evaluate the relationship of MCAS scores to outcomes from other teacher evaluations. If high MCAS scores are related to high scores on the other measure, that would provide some convergent evidence that MCAS scores are valid indicators of teacher effectiveness. Unfortunately, that is outside of the scope of this work and validity evidence based on relations to other variables will not be explored in this study.

Validity evidence based on consequences of testing refers to a very specific, and often confused, aspect of a test. Evaluating the consequences of testing involves a very specific question: are the decisions made based on the interpretation of test scores accurate? The accuracy is situation specific and must be evaluated for every proposed use of test scores. This means evaluating whether or not high scores on a mastery test identify students who will be successful in an Advanced Placement class; evaluating whether or not an employment test effectively identifies individuals for suitable positions; or evaluating whether or not a test identifies teachers who are diligently planning instructions following the guidelines of the content standards.

This category of validity evidence is often confused with consequences of testing related to graduation eligibility, “teaching to the test”, awarding or withholding merit pay, loss of employment, and many other examples. These consequences are a result of policies related to test scores, not the test itself. For example, there is validity evidence supporting the use of MCAS scores as indicators of student achievement. The decision to use MCAS scores to determine eligibility for graduation is purely a policy decision. A test may have ample validity evidence in place to support a proposed interpretation of test scores, yet the policies related to the test scores may be wildly inappropriate. The policies are not a function of the test and should not be considered validity evidence or used to build a validity argument (The Standards, 1999).

This study will attempt to address this category of validity evidence. First, though, we must remember that we are not evaluating whether or not MCAS scores should be used as indicators of teacher effectiveness. Their use as indicators of teacher effectiveness is purely a policy decision and thus not considered as validity evidence. What we must, and can, begin to evaluate is whether or not the decisions made based on this interpretation of test scores is accurate. This means, essentially, that there must be a link between student MCAS scores and the degree to which the teacher uses the Frameworks to plan instruction. Again, there must be a link between high MCAS scores and teachers who consistently use the Frameworks to plan instruction, lower MCAS scores and teachers who use the Frameworks to plan instruction to a lesser degree, and poor MCAS scores and teachers who do not use the Frameworks to plan instruction. To do this we can look at the relationship between the teachers' use of the Frameworks and the MCAS scores of their students.

These five broad categories should be the focus of the validity argument. The strength, or weakness, of the validity argument provides support for the proposed conclusions based on test scores. A strong validity argument will likely support the use of scores for a particular purpose, while a weak validity argument may indicate that the intended inferences and conclusions are not appropriate.

One of the basic tenets of a validity argument, which is implicit in any validity argument, is that the behavior captured on a test generalizes beyond the testing situation (Kane, 1990), thus the Standards' reference to validity generalizability. For example, we assume that performance at the time of testing in geometry can be generalized to represent performance on all occasions that geometry is called for. The generalizability of test scores to other situations is an issue that is still debated in the measurement community to this day. The intention of the Massachusetts DESE to now use MCAS results as a significant indicator of teacher effectiveness is taking a huge step away from the notion of validity as the term is traditionally used and as it is addressed in the Standards. We are now saying we can draw inferences, based on test scores, about people who are not even taking the test - validity one step removed.

This leap is justified by the reasoning explained earlier: valid interpretations about teacher effectiveness in teaching the Frameworks can be made based on MCAS scores because MCAS is aligned to the Curriculum Frameworks (content standards) and teachers teach the content in the Curriculum Frameworks; therefore, student achievement as measured by MCAS indicates how effectively a teacher teaches the Curriculum

Frameworks. This is the foundation for the proposed new use of MCAS scores and will be the focus of the validity argument.

There is any number of assumptions implicit in this argument. Several were noted in the introduction, and another researcher would likely posit many more. This study will begin to examine some of the underlying assumptions that are inherent in the line of reasoning presented above; specifically, that teachers teach the content in the Frameworks. The focus of the validity argument, for these purposes, will be on the degree to which Massachusetts teachers use the Curriculum Frameworks to plan instruction, the sensitivity of MCAS in detecting differences in teacher practice, and reasons why teachers may deviate from the Curriculum Frameworks. As you will recall, the Curriculum, Assessment, and Learning EEF Standard specifies that an effective teacher "promotes the learning and growth of all students through planning, instructional, and assessment activities that support a cycle of creating lessons focused on clear learning objectives…" (DESE, 2011a, p. 16). MCAS scores then fall into the evidentiary category of multiple measures of learning, growth, and achievement, which provide evidence of the degree to which the teacher fulfills the expectations of the Standard and is thus judged effective. The empirical evidence gathered during the course of constructing the validity argument will either provide support for or weaken the case to use MCAS scores as an indicator of teacher effectiveness by examining the underlying assumptions necessary to support the claim.

Instructional Alignment

The literature indicates that, while there are some individual teachers and some groups of researchers concerned with the congruence between instructional planning and content standards, the practice of deviating from the content standards is pervasive. As far back as the 1970s some educators were advocating the use of some kind of framework to ensure that the content of instruction and the content included on assessments were aligned (Steele, 1970, Cooley & Leinhardt, 1978; Popham, 1978; Leinhardt, 1981; Levine, 1982; Yalow & Popham, 1983; Oakes, 1986; Cohen, 1987; Winfield, 1993). There were also instances of litigation resulting from (perceived and legitimate) improper uses of assessment results (Yalow & Popham, 1983).

These early articles were based primarily on issues of equity and the increasing use of criterion-referenced testing. Criterion-referenced testing was an emerging concept in the 1970s, and there is a good deal of literature addressing "best practices." In an article advocating the merits of criterion-referenced tests, Popham (1978) called for the creation of achievement tests, tests that demonstrate what a student can do or has learned, that are criterion-referenced. Specifically, they must be linked to a well-defined behavioral domain (created by the test publisher in this article). Only in this way, the authors argue, can mismatch between test and instructional content be avoided and can students demonstrate mastery of curricula. Furthermore, Popham tells the reader, this is the only equitable method that allows decisions to be made based on inferences about the test scores. The concept of eliminating mismatch between test and instructional content gains momentum in the literature from that point, as demonstrated by the articles referenced above.

In the late 1990s and early 2000s some more focused attention was given to the notion that there must be alignment between content standards and instruction due to the advent of standards-based reform efforts in education, which came to the forefront of education with the implementation of NCLB. Researchers and educators became concerned that the validity of the interpretations of large-scale standardized assessment results would be weak if students were not taught the content being tested (Ananda, 2003; Anderson, 2003). The literature on this topic is surprisingly scarce. This paucity of research is possibly due to the fact that NCLB mandates proof of alignment between the content standards and the assessment, but does not *mandate* alignment between the content standards and instruction. The terms most commonly used for this concept are instructional alignment, curricular alignment or opportunity to learn.

The literature specific to instructional alignment describes studies that were very often conducted on a small scale and specific to a particular situation. Elia (1994) conducted a study of instructional alignment on a vocabulary test. The vocabulary words, and variants of the words, were taught to three separate groups, three different ways. Each group was then tested twice, one test aligned to the instruction and one test not aligned with instruction. Not surprisingly, the scores on tests aligned with instruction were found to be (statistically) significantly higher than those on the non-aligned test. Elia concluded that alignment of instruction with the test alone explained sixteen percent of the variance in the test scores.

Bober, Sullivan, Lowther, and Harrison (1998) undertook a study of the classroom practices of teachers enrolled in a master's level graduate program. There were five variables of interest in this study: learner-centered instruction; instructional design; media and technology; assessment; and instructional alignment. Bober, et al, define instructional alignment as "structuring the key components of the instructional process so that the instruction and the assessment are aligned with the instructional objectives" (p. 83). The authors developed a survey that was administered to the classroom teachers.

Bober, et al, only asked two survey questions, out of thirty, related to the degree to which objectives aligned with the assessment are considered in instructional planning. The results indicate that, on average, teachers "often" (as opposed to always or very often) consider the objectives that they will assess. Unfortunately, other than to report the mean response for these items, there is no discussion related to this finding.

Stein (2004) proposes a framework to evaluate the utility of commercially available mathematics curriculum for use in schools (Stein, 2004). Stein addresses instructional alignment in this article as the presence of a topic in the curriculum that also appears on the assessment. The framework is based on the Direct Instruction approach to teaching which applies specific principles of instructional design to curriculum development. Studies indicate that the Direct Instruction approach results in higher achievement than other instructional methods (Stein, 2004). The proposed framework for evaluating the adoption of mathematics curricula describes the importance of time allocated to the screening and evaluation of materials, the composition of committees who actually adopt the materials, and screening of the curricula under consideration. The framework then details procedures, a checklist type of approach, for evaluating the presence of a clearly defined content domain, the need to consider how the content will be incorporated in instruction, and the importance of the link between the content of instruction and the content on the assessment(s). The author advises that the use of this framework will lead to the thorough evaluation of commercially available mathematics curricula.

In 2004, Petersen and Cruz proposed a framework to use for planning physical education instruction that helps ensure alignment between learning objectives, teaching and learning activities, and assessments. They advise that the teacher have a clear focus on the point of the lesson. Teaching and activities should promote the focus of the lesson, provide meaningful feedback on performance, and be sure the closure (formal or informal assessment) is focused on the point of the lesson. The second step in their framework, lesson presentation, speaks to the type of instructional alignment of interest here. The authors stress that this focused approach to planning instruction can be the difference between students that are simply busy and students that are actually learning.

Smith (2008) proposed a framework for evaluating course design in higher education that focuses on the effectiveness of teacher practices in producing the desired outcomes (objectives). In this article, instructional alignment refers to the link between the processes employed by the teacher (e.g., teaching method, learning activities) and the outcome. This framework moves closer to the idea that there must be an explicit relationship between what is being taught and the learning objectives/outcome; and, although it was intended for higher education, it seems well-suited for adaption to the standards-based environment of public education.

James, Griffin, and Dodds (2008) were concerned about instructional alignment in physical education. The authors studied the learning objectives, instructional activities, and assessments of two physical education teachers. The two physical education teachers were videotaped, observed, and interviewed to determine the degree to which their instruction aligned with assessments. The authors concluded that neither teacher aligned their teaching/learning activities with their stated objectives (which can be thought of as content standards); therefore, the assessments aligned with their learning objectives were actually misaligned with the instruction.

Penuel, Fishman, Gallagher, Korbak, and Lopez-Prado (2008) report a study conducted in the state of Alabama, which looked at the implementation of a mandated science curriculum by teachers. The purpose of the study was to explore the ways in which teachers interpret state policies and implement the science curricula. Integral to the study was to understand the impact of professional development training about the new science curricula. Teachers who participated in a professional development training to learn about the new, policy-driven science curricula filled out a questionnaire asking about their experience implementing the new science curricula. The teachers were asked to think aloud while responding to the questions. Hierarchical linear modeling was used to explore the relationship between teachers' perceptions of alignment and alignment tools for the new curricula and the existing state content standards, interpretation of barriers to implementation, supportive experiences, the effect of adequate yearly progress on the school, and teachers' perceptions of the new science curricula.

The results of the study indicate that teachers implemented the science curriculum based on the availability of instructional material within their school, their perceptions of which standards the available material lent itself to teaching, and their own goals for instruction. Furthermore, the professional development the teachers received did not have the intended effect. It had little influence on the teachers' perception or actions regarding implementation of the new science curricula (Penuel, et al., 2008). The authors conclude that teachers perhaps have too much freedom in deciding which parts of curriculum are delivered to students, as well as how it is delivered. States are urged to continue to work toward better alignment by developing strategies that promote professional development and curriculum implementation strategies that are beneficial to teachers and all stakeholders in the system (Penuel, et al., 2008).

There are also some examples of studies investigating the variables that influence opportunity to learn. The importance of these variables is recognized as influencing students' ability to access the content standards. OTL is generally accepted to be the ability of schools to provide students with appropriate learning opportunities (Scherff & Piazza, 2009). This definition is broad in the sense that it relates not only to actual instruction on what will be tested, but also the resources available to the students. "Resources" is an ambiguous term in this literature, ranging from materials provided to the student to teachers' years of experience. The fundamental difference between instructional alignment and OTL is that OTL focuses on inputs (Scherff & Piazza, 2009), rather than outputs such as assessments in instructional alignment. This next section will provide a brief review of some of the recent OTL research.

These studies focus on teacher characteristics, such as education and years of experience (Boscardin, Aguirre-Munoz, Stoker, Kim, Kim, & Lee, 2005); resources made available by the school (Aguirre-Munoz & Boscardin, 2008; Scherff & Piazza, 2008); and the influence of textbooks on instruction (Tarr, Chavez, Reys, & Reys, 2006).

Tarr, Chavez, Reys, and Reys (2006) surveyed thirty-nine mathematics teachers in six states to ascertain the extent to which the textbook influenced instructional planning. The authors discovered that over half of the teachers surveyed used the textbook in approximately 90% of their instruction. The implication here is that textbooks are being used to plan instruction. This is an alarming discovery because textbooks are not (usually) aligned with state content standards; therefore, instruction is not likely to be aligned with state content standards. This misaligned instruction has the potential to deny students the opportunity to learn the state content standards upon which they will be assessed.

Scherff and Piazza (2008) studied OTL in terms of input resources for literacy programs. The resources of interest were content, curriculum activities, and materials. The authors were interested in students' perceptions of these resources, so they developed a survey for students. Over three thousand public school students in Florida, grades nine through twelve, responded to their survey.

Results indicate that there are three spheres of influence on student OTL: systems; offerings; and acknowledgement. Systems influences relate to issues such as withholding literary works from students in grades participating in high stakes testing and focusing on writing essays to prepare for the test. Essentially, the conclusion is that high stakes, standardized testing is negatively impacting OTL for the students in the survey. An offerings problem occurs when students only receive instruction in curricula that represents "test-like content" (p. 348) and when a textbook figures too prominently in instruction. The acknowledgement problem arises because students' voices are not included in determining how to meet educational goals. The findings from this study, which was designed to improve OTL in literacy programs, illustrates the complicated and many-faceted nature of OTL research. There are innumerable inputs influencing students each day and OTL has a direct impact on students' access to the content standards.

Boscardin, Aguirre-Munoz, Stoker, Kim, Kim, and Lee (2005) set out to investigate how certain OTL variables impact student performance and if the OTL variables have different effects in different subject areas. To that end, the authors examined the impact of OTL variables on English and algebra assessments. A teacher survey was used to determine OTL in the areas of teaching experience, teacher expertise in content area, content coverage, classroom activities, and assessment strategies and preparation. Socioeconomic status was also a variable of interest because a majority of the students completing the exams were from low socioeconomic backgrounds. 118 English teachers and 4,715 of their students and 124 algebra teachers and 4,724 of their students participated in this study.

Results show that three variables were significantly related to student performance on the English and algebra assessments: Socioeconomic status; teacher expertise; and content coverage. Basically, as teacher expertise and content coverage increased, and the number of low socioeconomic students decreased, test scores increased. The analyses reported were conducted using a two-level hierarchical linear model.

The findings regarding content coverage and socioeconomic status on test performance are not surprising and simply confirm the results of similar studies (Boscardin, Aguirre-Munoz, et al, 2005). What is significant here is the relationship between teacher expertise and student performance. Inner city schools do not generally attract the best or most experienced teachers, yet the majority of students in inner city schools come from a low socioeconomic status background. The authors look at this as an OTL variable because the expertise of the teacher directly impacts their OTL the subject.

The first two authors of the study summarized above, Aguirre-Munoz and Boscardin (2008), contend that OTL has taken on an increasingly important role in education post-NCLB given the disparate distribution of educational resources to students. The authors believe a measure of OTL should be developed. Furthermore, the impact of OTL, as indicated by some measure, should be used to interpret test scores because there is evidence to suggest that OTL variables explain test scores (Aguirre-Munoz & Boscardin, 2008).

An OTL survey was completed by twenty-seven teachers and a language arts assessment was given to over one thousand students. The relationship between OTL, via the teacher survey, and student scores on the assessment was studied. The teacher survey was designed to capture six critical aspects of OTL in the classroom: Teaching experience; teacher expertise in content topics; content coverage; classroom processes; assessment practices and assessment preparation; and classroom resources. An ordinal logistic hierarchical modeling analysis indicated that only two of the OTL variables had significant effects on student performance on the assessment. The variables were literary analysis and writing. Essentially, the amount of instructional time spent on literary analysis and writing influenced these test scores (Aguirre-Munoz & Boscardin, 2008).

The articles summarized above have a common theme in that they recognize the importance of testing what has been taught. There are concerns about the impact on fairness, bias, and validity when alignment between test content and instructional content is ignored. The remainder of the articles included in this literature review will focus on research that attempts to develop methodologies and instruments designed to assess the alignment between content standards and instructional content, as well as post-NCLB attitudes toward content standards and instructional content alignment.

Proposed alignment frameworks

An article from Anderson (2002) reveals that some researchers were concerned with the limited scope of the term alignment from nearly the inception of NCLB. Anderson (2002) clearly delineates the three elements of an assessment system that must align: Content standards; instruction (including materials); and outcome measures. Anderson (2002) represents the relationship between the components with a triangle with one of the three elements on each point. The relationship represented by each arm of the triangle is equally important, according to Anderson, and must be attended to in alignment studies. Anderson calls the alignment between the three elements curricular alignment (please see Figure 1).

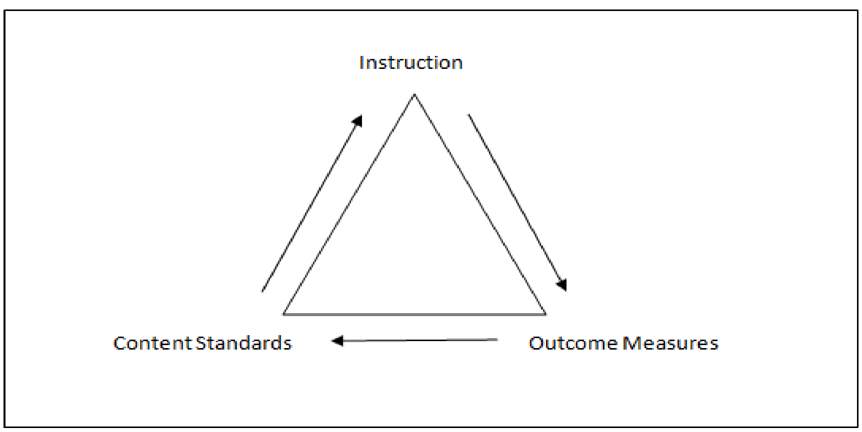

Figure 1. Anderson's Curricular Alignment.

Anderson (2002) proposes a framework to help makes sense of data collected in a curricular alignment study for ease of interpretation. The proposed framework is called the Taxonomy Table. The Taxonomy Table, a revision of Bloom's, is intended to facilitate estimating alignment for any subject at any grade level, spotlight student learning, and provide reasonable validity estimates of alignment (Anderson, 2002). There are four steps in this process: Objectives (content standards) are put in the appropriate cells in the table; instructional activities are placed in their cells of the table; each assessment task is placed in a cell in the table; and, finally, comparisons are made between the Taxonomy Tables for each element of curricular alignment.

Evidence of alignment is provided when objectives from the content standards, instructional activities, and items from assessments fall into the same cells on their respective Taxonomy Tables. Please refer to Table 2 for an example (Table 2 reproduced from Airasian & Miranda, 2002).

Table 2. The Cognitive Process Dimension.

The Knowledge Dimension	**1. Remember**	**2. Understand**	**3. Apply**	**4. Analyze**	**5. Evaluate**	**6 Create**
A. Factual Knowledge	**Objective 1** Days 2, 3, & 5 Activities *Assessment B*					**Objective 3** Days 8-10 Activities *Assessment C*
B. Conceptual Knowledge		**Objective 2** Days 1, 4-7 Activities *Assessment A*		Days 6-7 Activities	**Objective 4** Days 8-10 Activities	**Objective 3** Days 8-10 Activities *Assessment C*
C. Procedural Knowledge			Day 4 Activities *Assessment C*			
D. Metacognitive Knowledge						

Key
Objective 1: Remember the specific parts of the Parliamentary Acts.
Objective 2: Explain the consequences of the Parliamentary Acts for different colonial groups.
Objective 3: Choose a colonial character or group and write a persuasive editorial stating his/her/its position on the Acts.
Objective 4: Self- and peer-edit the editorial.
Assessment A: Classroom questions and informal observations
Assessment B: Quiz
Assessment C: Performance Assessment (editorial, with 10 evaluation criteria)

Table 2 is an example of relatively strong alignment between objectives (akin to content standards), instruction, and assessment. Cells A1, A6, B2, and B6 explicitly link instructional activities and assessment to an objective. Partial alignment may occur when the elements fall into the same column, but different rows, or vice versa (Anderson, 2002).

The framework described here also provides evidence of partial alignment, which can be seen within Table 2. Cell B5 shows instructional activities tied to an objective, but with no assessment planned. Cell C3 shows instructional activities and an assessment that are not linked to an objective. This information can also be helpful because partial alignment has the potential to provide diagnostic information as it illustrates mismatch between the components. Teachers, for example, can use this information to adjust instruction so it more closely aligns with the content standards and assessment (Anderson, 2002).

In his 2002 Presidential Address in Educational Researcher, Porter calls for an expanded understanding of alignment. Porter (2002) includes the "content of instruction, educational materials, content standards, and professional development" (p. 3) as key components in educational reform. Porter (2002) opines that understanding the role of the content of instruction is vital to understanding student achievement within standards-based reform programs. He views content of instruction as a variable that affects student achievement and thus warrants careful research. Porter (2002) advocates alignment methods that allow researchers to gauge the effect of content of instruction on student achievement.

Porter's (2002) views on the matter are no surprise as, in 2001, he and Smithson developed an alignment methodology that facilitated a quantitative comparison of alignment between content standards, instruction, and assessment across teachers, schools, and states. The methodology is called Surveys of Enacted Curriculum (SEC) and was used to evaluate the link between content standards, instruction, and assessment in 11 states and four urban school districts (Porter & Smithson, 2001).

The SEC uses a matrix design to measure the degree of alignment between content standards, instruction, and assessment (Porter & Smithson, 2001; Porter, 2002; Porter, Smithson, Blank, & Zeidner, 2007). The incorporation of an instructional content component in the SEC distinguishes this method from many others that focus only on the link between content standards and assessment. This feature is particularly important when evaluating student achievement in this era of standards-based reform, because instructional content is an intervening variable (Martone & Sireci, 2009; Porter, 2002). The same matrix design can be used for alignment studies between any combination of content standards, instruction, and assessment; it also allows for content analysis of instructional materials and assessments.

Of primary concern for the SEC is the development and use of a uniform language to describe topics and the cognitive demand (levels of thinking) placed on the student. The uniform language forms the foundation for curriculum indicators which guide data placement in the matrix. Teachers, who supply the data, must be trained in the proper use of the language of the SEC to properly code their instructional content for data analysis.

Daily teacher logs, observation protocols, or specially developed teacher surveys are used to gather the necessary data for the matrix. It should be noted that the teacher surveys, in an effort to reduce the complexity of coding and avoid placing an undue burden on the participating teachers, reduce the number of cognitive demand categories. This is a bit troublesome as daily teacher logs and observation protocols are time and cost prohibitive, resulting in the use of the surveys in pilot studies (Martone & Sireci, 2009). Regardless of the data collection method, the purpose is to explore instructional content, the amount of time teachers spent on a topic, and to determine which cognitive demand categories were emphasized. Survey responses are reported on a Likert-type scale.

Once the curriculum indicators are coded, the data are transformed into proportions representing the total amount of instructional time dedicated to each cell in the matrix. The proportions from the cells can then be used to calculate an index of alignment between matrices. The index of alignment is calculated as follows (Porter, 2002):

$$1 - \frac{\Sigma\left|X - Y\right|}{2}$$

where X denotes cell proportions from one matrix and Y denotes cell proportions from another matrix. The two matrices of interest for my purposes are content standards and instructional content; however, you will recall that the SEC matrices and indices of alignment could be calculated for any two components of the content standards-instruction-assessment-achievement cycle. The values range from 0 to 1.0. Perfect alignment is seen with a value of 1.0. Porter (2002) also suggests that correlations can be calculated across cells between two matrices to examine the relationship between content proportions.

The SEC results can be graphically displayed on a content map which is laid out much like a topographical map. Once the proportions of time spent on instruction of each topic are calculated, the authors suggest creating content maps to graphically display the emphasis of topic by cognitive demand. This allows direct comparison of content maps for different components of interest. The graphical display shows the areas of overlap between content standards and instructional content, allowing researchers to make judgments about alignment.

Content maps are necessary to compare SEC results because the index of alignment is interpretable only relative to other indices of alignment (by comparison). That is to say, a larger value on the index of alignment is obviously better, yet there is no way to know what value demonstrates adequate alignment. The reader is also advised to think of interpreting the values of the alignment index normatively (Porter, et al, 2007). For example, if the alignment index value for one state is higher than the average alignment index values for other states, the conclusion is that alignment for that one state is high. The authors advise that the alignment index is particularly useful for making comparisons between teachers (for instructional content coverage), schools, states, content standards, assessments, or “anything else that can be content analyzed” (Porter, et al., 2007, p. 46).

Porter, Smithson, Blank, and Zeidner (2007) revisited the SEC alignment matrix methodology with the intention of expanding on its key features, one feature being the quantitative index of alignment. The matrix is used to calculate the proportion of content devoted to a topic, as previously described. The authors introduce the idea of performing analyses based on smaller blocks of time than the original school year described in the first article, because teachers will be better able to describe their instruction for smaller blocks of time. The results of multiple analyses are to be aggregated to the school year (Porter, et al., 2007).

Teacher practice

The literature and frameworks reviewed here are examples of decades-old concern with ensuring that students are taught the content standards for reasons of fairness, but also because it directly impacts the validity of the interpretations of test scores. To extend this reasoning, if we are to judge the effectiveness of a teacher against a criterion, the content standards as measured by MCAS, we must be sure that the criterion is an accurate representation of what teachers are teaching, much the same as we need to know that the content standards are an accurate representation of what students are being taught. There is a body of research to suggest that many teachers deviate from teaching the content standards. A review of the literature concerning teachers' use of the content standards to plan instruction follows.

In 2001, Jacob reported that teachers have been advised for years to align instruction with standards. This sentiment is echoed by Porter (2002), and taken even further with his assertion that "the content of instruction plays a primary role in determining gains in student achievement" (p. 3).

In his article, Porter acknowledges that content standards pass through the lenses of individual teachers and, therefore, the content delivered and the manner in which it is delivered will vary as teachers vary. Porter recognizes that the experiences of the teacher (as both student and teacher), teacher preparation, familiarity with a subject, and the materials available to each teacher will dictate the content of instruction. His point, which may seem counterintuitive given that content standards are at the root of what should be taught, is echoed by other authors (Porter, 2002; Remillard & Bryans, 2004; Sherin & Drake, 2010). Sherin and Drake (2010) report that some teachers change the content of reform-based materials (such as content standards), either consciously or unconsciously, effectively bypassing the goals of reform.

There is little research in the education literature addressing teachers' use of content standards in instructional practices; however, in 2000, Education Week reported results of the Reality Check survey. The Reality Check survey is a survey of attitudes and practices in education. A random, nationally representative sample of approximately 2,300 students, teachers, parents, professors, and employers were surveyed. One very interesting finding from the survey comes from the 604 public school teachers interviewed. Content standards were in place for the vast majority of responding teachers (97%), yet only 42% of teachers reported receiving "most of the guidance about what they should teach from state standards" (p. 4). Reality Check results demonstrate that despite the changes related to public education reform, the teaching practices of some have failed to change.

These findings are supported, albeit implicitly, by a report from the American Federation of Teachers (AFT) published in 2009. The report, titled The Instructional Demands of Standards-Based Reform, compares traditional instructional planning methods with methods required to effectively plan instruction based on curriculum standards. The AFT report calls for teachers to begin lesson planning by considering the standard(s) being taught and focusing on what types of evidence will demonstrate that the standard has been learned. This is in direct contrast to traditional lesson planning in which individual teachers determine the topic, plan a lesson and activities, then develop an assessment to determine how much has been learned and by whom.

The AFT report points out that teachers are “woefully unprepared” (p. 5) to plan instruction as required in a standards-based system. The report states that simply instituting a standards-based system does not mean improved achievement will follow because this type of system requires behaviors of teachers (and students) that have never before been necessary. A similar observation is made in a report from Hamilton, Stecher, and Yuan (2009). The authors point out that although standards-based reform efforts (e.g., content standards) influence practice, they do not change fundamental pedagogical beliefs. As teachers continue to exercise a high degree of autonomy in what and how they teach, there is a lack of consistency in teachers’ educational practices, including instruction. This can eventually result in a conflict wherein autonomy and alignment with the content standards become competing goals for teachers (Hamilton, et al, 2009). The AFT report stresses the need for teacher preparation courses, both pre- and in-service, and professional development that teaches effective instructional planning within a standards-based system.

Recognizing the need to assist teachers in effectively incorporating the standards in lesson planning in the era of standards-based education reform, programs have been developed that are intended to support teachers with planning instruction in a standards-based system (Childre, Sands, & Pope, 2009). Some of these programs, using the term loosely because it is not necessarily a formal process or even a requirement, are based on the use of a specific framework for planning lessons, such as Understanding by Design or backward design (Wiggins & McTighe, 2005). This is a form of curriculum mapping wherein every lesson is explicitly linked to a standard or multiple standards from the Curriculum Frameworks. Curriculum mapping is a system that keeps track of the skill being taught (the standard), the actual content of the lesson intended to teach the skill, and what type of assessment was used to evaluate whether or not the student learned the skill (Jacobs, 2000). Backwards design is an extension or modification of curriculum mapping which requires the teacher to develop an assessment for the lesson that considers what evidence will demonstrate learning *before* they develop the actual content of the lesson. The goal is to reduce the practice of planning a lesson, teaching the lesson, then putting together an assessment which results in a grade for the grade book. Such practice may (or may not) have been effective at one time, but it is not an effective approach to teaching in a standards-based education system according to Wiggins and McTighe (2005).

Other programs implemented by schools across the country include the installation of a curriculum director, whose sole job is to act as a bridge between the content standards and teachers. Curriculum directors may, depending on the school, decide which standards will be addressed, decide what material (e.g., handouts) will aid in teaching the standard and provide the material to the teacher, decide the timeline of teaching the standards, and decide how much time is to be spent addressing a standard (Hamm, 1994). The curriculum director is also often responsible for evaluating programs proposed for use in the school, such as reading or math programs, and evaluating professional development opportunities for teachers. The work of the curriculum director is not focused only on the content standards (Hamm, 1994). Some schools use curriculum pacing requirements (Tennessee Department of Education, 2011) which are guides based on the academic calendar that detail which content standards are to be taught, the order they are to be taught in, and the amount of time to be given to each standard. In Tennessee they are developed either by teams of teachers from the same grade or individual teachers. And some schools leave instructional planning, thus instruction of the content standards, solely to the discretion of individual teachers (Childre, 2009). The variable range of practices used to assist teachers in planning instruction supports the research literature's conclusions that teachers are not prepared to plan instruction in a standards-based educational system.

These varying levels of teacher preparation, as well as the implementation of different practices in pursuit of curricular (or instructional) alignment, likely result in a wide range of teacher practice in planning instruction. The validity of interpreting MCAS scores as an indicator of teacher effectiveness in teaching the Frameworks requires that such differences are reflected in the scores attained by the students. Furthermore, it also requires that the scores are sensitive enough to the differences to allow evaluators to separate teachers into different categories of effectiveness. Given the high pass rate of students taking the MCAS, strictly focused on the score of 220 as opposed to Proficient and Advanced, it may be difficult to defend the use of MCAS scores as a measure of teacher effectiveness in teaching the Frameworks. In order to investigate these concerns, research into teacher practices related to the use of the Frameworks to plan instruction and the sensitivity of the MCAS to differences in teacher practice is necessary.

Research Questions

The purpose of this research is to examine some of the assumptions underlying the use of MCAS scores as a significant indicator of teacher effectiveness. In accordance with the Standard's recommendation to construct a validation argument for each proposed use of test scores, we have identified some of the propositions inherent in the use of MCAS scores as an indicator of teacher effectiveness. The starting point for building a validation argument supporting the use of MCAS scores as an indicator of teacher effectiveness will be the following rationale: Valid interpretations about teacher effectiveness in teaching the Frameworks can be made based on MCAS scores because MCAS is aligned to the Frameworks and teachers teach the content in the Curriculum Frameworks; therefore, student achievement as measured by MCAS indicates how effectively a teacher teaches the Frameworks.

A logic model is a graphical representation of a process that explicates the assumptions underlying a situation that lead to a particular result (Millar, Simeone, & Carnevale, 2001). Basically, logic models are intended to make it easier to understand a process. Figure 2, presented below, is a logic model that attempts to make it easier to understand the relationship between the different aspects influencing the validity argument for using MCAS scores as indicators of teacher effectiveness.

SITUATION	INPUTS	ACTIVITIES	OUTPUTS	OUTCOMES	
				Knowledge	Actions
The logic for using MCAS scores as a measure of teacher effectiveness is that since we know what teachers are supposed to teach, we can measure the students to see if they have learned it.	Teachers plan instruction using the Frameworks.	Teachers teach the contents of the Frameworks. Students take MCAS.	Student performance on MCAS (MCAS scores).	Degree of teacher effectiveness.	Teachers classified into performance level. Teachers receive rewards/incentives or sanctions based on the performance level classification.

ASSUMPTIONS

Teacher preparation programs prepare the students to plan instruction based on content standards.
Alignment of MA Curriculum Frameworks and MCAS.
MCAS scores are sensitive enough to reflect variations in teacher practice.

EXTERNAL FACTORS
Depndant on student performance.
Lack of research supporting this practice.

Figure 2. Logic Model.

As you can see in the logic model, some of the propositions, or assumptions, underlying this validity argument, and the propositions that will be investigated here, are that Massachusetts teachers teach the content of the Frameworks; MCAS scores reflect the degree to which a teacher teaches the Frameworks (e.g., a high student score indicates the teacher has covered the Frameworks well); and MCAS scores are sensitive enough to variations in teacher practice (if they exist) to indicate which performance category a teacher should be assigned to. Therefore, the research questions are:

1. Do Massachusetts teachers teach the content of the Frameworks?

    1a. Are there variations in teacher practice related to the use of the Math Frameworks for planning and implementing classroom instruction?

    1b. Are teacher behaviors related to the use of the Math Frameworks associated with gender or grade taught?

    1c. Is a teachers' years of experience associated with teacher practice related to use of the Math Frameworks?

    1d. Are teachers' opinions about autonomy associated with teacher practice related to the use of the Math Frameworks?

    1e. Are teacher preparation programs associated with teacher practice related to the use of the Math Frameworks?

    1f. Is participation in professional development designed to assist teachers in using the Math Frameworks to plan instruction associated with the way teachers use the Frameworks?

    1g. Are teachers' opinions toward the Math Frameworks associated with use of the Math Frameworks to plan instruction?

2. If variations in teacher practice related to the use of the Frameworks exist, is the MCAS sensitive to these variations such that they are reflected in student performance?

2a.What is the relationship between the teachers' self-reported use of the Frameworks and student performance, as reported by the respondent?

2b. What is the relationship between a principal's attitude toward the Frameworks and MCAS scores for the school?

2c. What is the relationship between teachers' use of the Frameworks and MCAS Math scores, by grade level?

The information gathered in the course of investigating these research questions will be used to inform a validation argument pertaining to the use of MCAS scores as a measure of teacher effectiveness. If the propositions underlying the validity argument do not hold, it will raise serious questions about the validity of MCAS scores as an indicator of teacher effectiveness.

CHAPTER 3

METHOD

In order to investigate the proposed research questions, an evaluation of teacher practice is essential. This will be accomplished via surveys of Massachusetts public school teachers and principals. The idea of using the SEC, described previously, was considered and rejected due to the time involved in an SEC alignment study. The SEC requires that the participants be trained to use a language developed specifically for the surveys and that many lessons are evaluated in order to develop the content maps that allow for meaningful interpretation of the results. Given the time-consuming requirements placed upon participating teachers, it was considered unlikely that an adequate sample of teachers would volunteer to participate. Survey instruments were developed for this study. Before embarking on the full study, a pilot study was conducted to determine the extent to which teachers are using the Curriculum Frameworks in Massachusetts. Although the literature suggests that many teachers do not use the Frameworks, those studies do not necessarily generalize to Massachusetts teachers.

Pilot Study

In light of the preceding evidence that some teachers do not plan instruction that is aligned with the curriculum standards developed or endorsed by their state government, an exploratory survey to assess Massachusetts teachers' use of the Massachusetts Curriculum Frameworks in planning instruction was developed. The survey asked questions related to teachers' familiarity with, use of, and attitude towards the Massachusetts Curriculum Frameworks. Teachers were also asked if they would like help using the Curriculum Frameworks to plan instruction and, if so, in what form. This survey was intended to be a preliminary, or exploratory, tool used to guide further research into the issue of instructional alignment. Of particular interest were teachers' attitudes toward the Curriculum Frameworks and instruction.

Method. A convenience sample of 24 teachers was surveyed at a Professional Development seminar in July of 2010. The seminar was developed to expose teachers to various forms of computer-based technology currently being used in schools and classrooms. The teachers either chose to attend the Professional Development themselves or were asked to attend by administrators in their school. Participants received $300.00 and Professional Development points for their involvement in the seminar. This is not a representative sample of Massachusetts teachers.

The survey, titled "Exploring the Relationship Between Teachers and Frameworks," was developed to address areas of concern identified in a review of the literature on this topic. Questions were also included based on observations of teachers' comments made by the author in the course of her professional and educational experiences. The survey questions were reviewed by two Professors of Education, one of whom is an expert in measurement and the other in curriculum development. Their comments and suggestions were used to guide revisions to the original survey and develop the final version used in this pilot study.

The survey contained 31 questions. There were six demographic type questions pertaining to years of experience and the type of school the teacher worked in, 14 Likert-type questions, 2 Yes/No questions, 2 "select all that apply" type questions, and 7 open-response questions. The survey explored three broad topics: attitude towards and use of

Curriculum Frameworks, attitude towards and use of MCAS peripheral materials, and teachers' perception of support from administration.

The surveys were completed at the end of the two-day Professional Development seminar. Participation in completing the survey was voluntary. Respondents were informed that their answers would be anonymous, and asked not to identify themselves on the survey form. In an effort to elicit answers not colored by social desirability, respondents were assured information would not be reported in any manner that could connect responses to an individual respondent. Please refer to Appendix A to review the complete survey.

Full Study

Survey construction. The results of the pilot study presented in Chapter 3 indicate that some Massachusetts teachers do not use the Frameworks to plan instruction and others follow the Frameworks to a greater or lesser degree. These findings warrant further investigation into the practices of Massachusetts teachers' use of the Frameworks and how this may affect the use of MCAS scores as indicators of teacher effectiveness.

Based on the results of the initial survey and on issues raised in the literature review, a revised survey was created that attempted to gain a more in-depth understanding of how Massachusetts teachers use the Frameworks to plan instruction, their attitudes toward the Frameworks, and why teachers may deviate from the Frameworks when planning instruction. The focus of the surveys was shifted, however, to grades 3- 8 math teachers. The reason for this shift is that MCAS results will be used in the form of SGPs, thus the practices of teachers in these grades are most important because SGPs will be calculated based on their students' MCAS scores. Math became the focus because it was decided that the Math Frameworks left less room for variation in instruction than the English Language Arts Frameworks.

Questions were added to the survey related to teacher preparation and professional development in light of the research positing that teachers may be unprepared to plan instruction based on content standards because teacher preparation and professional development programs have failed to adjust their curricula in response to reform (Porter, 2002; AFT, 2009). There were also questions added to the survey related to what program, if any, is in place to help teachers use the Frameworks to plan instruction (Hamm, 1994, Jacobs, 2000; Wiggins & McTighe, 2006); and the teachers' opinions regarding the level of autonomy they should have in planning instruction (Hamilton, Stecher, & Yuan, 2009).

A survey of Massachusetts public school principals was also developed. The purpose of the Principal Survey is two-fold: it will help illustrate any discrepancies between school policy and teacher practice and it will act as verification of the information provided by teachers. For example, school policy may mandate teachers plan instruction in accordance with some State approved program (backward design, for example), but teachers may not adhere to the policy for various reasons. There are various implications that could arise from this information and that would need to be addressed when evaluating the interpretation of MCAS scores as an indicator of teacher effectiveness.

It was decided, for both the Teacher Survey and the Principal Survey, to focus the questions on the state Frameworks despite the recent adoption of the Common Core State Standards (CCSS). The State is planning a multiyear adoption and implementation process for the CCSS and not all school districts, or schools within a district, have adopted the new standards; however, all schools and districts have recently used or are currently using the Frameworks.

Sample. The participants were public school math teachers and principals from the state of Massachusetts. An email was sent to public school Teachers and Principals in the State of Massachusetts, in schools that included grades 3 -8, for whom an email address was available online, either via DESE, a district, or school website. The email introduced the researcher and described the purpose of the research. The recipients of the emails were also informed of an incentive, in the form of ten $100.00 Amazon.com gift cards that were to be awarded to ten randomly selected survey participants. The email contained a direct link to either the Principal survey or the Teacher survey. Participants clicked the link and were routed to the survey in Survey Monkey. Survey Monkey is an internet-based survey service where individuals can create, distribute, and collect responses to surveys.

Introductory emails containing a link to the survey were sent to 8,332 Massachusetts public school teachers. Please refer to Appendix B for a copy of the email. One week later another email was sent reminding the recipient of the purpose of the survey and the incentive. Please refer to Appendix D for a copy of the follow-up email sent to teachers. It was not possible to calculate a response rate for this sample because many of the recipients likely did not meet the specified requirement of teaching grades 3-8 Math. Unfortunately, very few school websites provided any information about the teachers other than a name and email address; therefore, it was often impossible to determine the grade or subject a teacher taught. Emails were sent to any teacher with an available email address in hopes of maximizing the number of grades 3 – 8 math teachers contacted. Responses were collected from 745 Massachusetts public school, grades 3 – 8 math teachers.

Introductory emails containing a link to the survey were sent to 1,662 Massachusetts public school principals. One week later another email was sent reminding the recipient of the purpose of the survey and the incentive. Please refer to Appendix C for a copy of the introductory email and Appendix E for a copy of the follow-up email sent to principals. Responses were collected from 147 Massachusetts public school principals. The response rate was approximately 9%.

Participation in the survey was strictly voluntary and thus motives for responding to the survey are very much unique to the individual respondents. As a result, this survey may not have resulted in a representative sample; therefore, results from this sample may not be generalizable to the larger body of Massachusetts math teachers, teachers of other subjects, or principals.

Surveys. The teacher survey, titled “Exploring the Relationship Between Teachers and the Frameworks”, was developed to address areas of concern identified in a review of the literature on this topic. Questions were also included based on observations of teachers’ comments made by the author in the course of her professional and educational experiences. The survey questions were reviewed by a Professor of Education who is an expert in measurement, a teacher currently employed in the Massachusetts public school system, and a public school Curriculum Director with a background in educational measurement. Their comments and suggestions were used to guide revisions to the original survey and develop the final version used in this study. A small pilot study of the live survey on Survey Monkey was also conducted to evaluate the appearance of the questions on the screen, the routing rules embedded within the survey, and the data collection procedure.

The Teacher survey contained 56 questions. There were 5 demographic type questions pertaining to years of experience, the type of school the teacher worked in, and what subject and grade the teacher taught; 42 five-category agreement scale questions; 3 questions asking teachers to provide information about student performance on MCAS; 3 Yes/No questions; 1 selected response question; and 2 open response questions. The survey explored four broad topics: attitude towards and use of Curriculum Frameworks, teacher preparation and professional development experiences, teachers' opinions of what would be helpful in assisting them to use the Math Frameworks to plan instruction, and opinions of the evaluation process. The survey was accessible through the link in the email for a two-week period in June of 2012. Please refer to Appendix F for a copy of the Teacher Survey.

The Principal survey, entitled "Exploring the Relationship between Principals and Teacher Evaluation", was developed to address issues related to teacher practice identified in the literature and to provide the administration perspective on teacher practices within the school. The survey questions were reviewed by a Professor of Education who is an expert in measurement, one Massachusetts public school teacher, and one public school Curriculum Director with a background in educational measurement. Their comments and suggestions were used to guide revisions to the original survey and develop the final version used in this study.

The Principal Survey contained 39 questions. There were 6 demographic questions, 30 five-category agreement scale questions, 1 yes/no question, and 2 open response questions. The survey explored four broad categories related to attitude toward the Frameworks, how teachers in the school use the Frameworks to plan instruction, what would be helpful in assisting teachers to use the Frameworks to plan instructions, and teacher evaluation. The survey was accessible through the link in the email for a two-week period in June of 2012. Please refer to Appendix G for a copy of the Principal Survey.

Analyses. The Teacher Survey and the Principal Survey were analyzed using SPSS (SPSS, Inc, 2008), a commercially available statistical software package. Descriptive statistics were calculated for all responses and responses to open-ended questions were analyzed and summarized by the researcher.

Many survey questions were used to measure teachers' attitude towards and opinions of the Frameworks, as well as the way teachers use the Frameworks to plan instruction. To allow for meaningful analyses, appropriate composite scores were created for each construct rather than conducting multiple analyses on individual survey questions. An exploratory factor analysis, a statistical procedure that models the relationship between the factors underlying questions on a test or survey (Kane, 2006), was performed on the Teacher Survey questions to develop composite score scales. Principal Factor Analysis was used to determine how many factors were represented by the subsets of survey items. Varimax rotation was used to find an interpretable solution. Information from the EFA, along with substantive interpretation of the factors underlying the survey questions, was used to identify items that were measuring the same factor. The items were combined to form subscale scores. Coefficient alpha was used to determine the reliability of each subscale. There were not a sufficient number of responses to the Principal Survey to perform exploratory factor analysis.

The specific research questions put forth in Chapter 2 were addressed as follows:

Descriptive statistics. Descriptive statistics were used to answer Research Question 1a. Descriptive statistics are used to summarize and present data (Gravetter & Wallnau, 2008), making it easier to interpret results. Specifically, responses to survey questions related to how often teachers use the Math Frameworks to plan instruction, how closely they follow the Math Frameworks, how often topics that are not included in the Math Frameworks are included in instruction, and how often teachers use a program that links the Math Frameworks with instruction to plan instruction were examined and responses compared.

- 1a. Are there variations in teacher practice related to the use of the Math Frameworks for planning and implementing classroom instruction?

Pearson Product Moment correlations. The Pearson Product Moment Correlation measures the strength and direction of a linear relationship (Gravetter & Wallnau, 2008). The Pearson Product Moment Correlation was used to address Research Questions 1c, 1d, 1e, 1f, and 1g. These research questions address the relationship between years of teaching experience, teachers' opinions about autonomy, differences in preparation programs, participation in professional development, and teachers' attitude toward the Math Frameworks and their use of the Math Frameworks to plan instruction.

- 1c. Is a teachers' years of experience associated with teacher practice related to use of the Math Frameworks?

- 1d. Are teachers' opinions about autonomy associated with teacher practice related to the use of the Math Frameworks?
- 1e. Are teacher preparation programs associated with teacher practice related to the use of the Math Frameworks?
- 1f. Is participation in professional development designed to assist teachers in using the Math Frameworks to plan instruction associated with the way teachers use the Frameworks?
- 1g. Are teacher attitudes toward the Math Frameworks associated with the use of the Math Frameworks to plan instruction?

Analysis of Variance (ANOVA). An ANOVA was used to investigate Research Question 1b. ANOVA evaluates the mean differences between two or more populations via hypothesis testing (Gravetter & Wallnau, 2008). The influence of gender and the grade taught on the degree to which the Math Frameworks are used to plan instruction were analyzed using an ANOVA.

- 1b. Are teacher behaviors related to the Math Frameworks associated with gender or grade taught?

The Use Scale was the dependent variable in the ANOVA. Gender and grade taught were the independent variables. A Tukey's post hoc test was used to make pairwise comparisons of the means of the Use Scale by grade level. Tukey's identifies significant mean differences between multiple groups while controlling the Type I error rate (Gravetter & Wallnau, 2008).

Linear regression. Linear regression is a statistical analysis that models the relationship between variables by finding the best fitting line for a set of data. This line, the regression line, allows you to predict values of one variable based on values from another variable (Gravetter & Wallnau, 2008). Linear regression was used to investigate Research Questions 2a, 2b, and 2c.

2a.What is the relationship between the teachers' self-reported use of the Frameworks and student performance, as reported by the respondent? A series of linear regressions was used to evaluate the relationship between teachers' self-reported use of the Frameworks and student performance, as reported by the respondent. In each of the linear regressions, the Use Scale scores were the independent variable and the percent of students scoring in the Needs Improvement, Proficient, and Advanced performance categories, respectively, was the dependent variable. The regression equations for each model were:

Needs Improvement model: Y = $\beta_0 + \beta_1 x_{NI}$

Proficient model: Y = $\beta_0 + \beta_1 x_P$

Advanced model: Y = $\beta_0 + \beta_1 x_A$

2b. What is the relationship between a principal's attitude toward the Frameworks and MCAS scores for the school? A series of linear regressions was used to evaluate the relationship between the principals' attitude towards the Frameworks and MCAS scores for the school. In each of the linear regressions, the principals' response to survey question 9, "The Math Frameworks are a good representation of the material that should be taught," was the independent variable. The percent of students scoring in the Needs Improvement, Proficient, and Advanced performance categories across the entire school, reported by DESE, was the dependent variable. The regression equations for each model were:

Needs Improvement model: Y = $\beta_0 + \beta_1 x_{NI}$

Proficient model: Y = $\beta_0 + \beta_1 x_P$

Advanced model: Y = $\beta_0 + \beta_1 x_A$

2c. What is the relationship between teachers' use of the Frameworks and MCAS Math scores, by grade level? A series of linear regressions was used to evaluate the relationship between teachers' self-reported use of the Frameworks and student performance, as reported by DESE. In each of the linear regressions, the Use Scale scores were the independent variable and the percent of students scoring in the Needs Improvement, Proficient, and Advanced performance categories respectively, was the dependent variable. Due to the manner in which DESE reports student performance data, the percent of students in each performance category was aggregated across all teachers for the grade level in each school. The regression equations for each model were:

Third Grade Needs Improvement model: Y = $\beta_0 + \beta_1 x_{3NI}$
Third Grade Proficient model: Y = $\beta_0 + \beta_1 x_{3P}$
Third Grade Advanced model: Y = $\beta_0 + \beta_1 x_{3A}$

Fourth Grade Needs Improvement model: Y = $\beta_0 + \beta_1 x_{4NI}$
Fourth Grade Proficient model: Y = $\beta_0 + \beta_1 x_{4P}$
Fourth Grade Advanced model: Y = $\beta_0 + \beta_1 x_{4A}$

Fifth Grade Needs Improvement model: Y = $\beta_0 + \beta_1 x_{5NI}$
Fifth Grade Proficient model: Y = $\beta_0 + \beta_1 x_{5P}$
Fifth Grade Advanced model: Y = $\beta_0 + \beta_1 x_{5A}$

Sixth Grade Needs Improvement model: Y = $\beta_0 + \beta_1 x_{6NI}$
Sixth Grade Proficient model: Y = $\beta_0 + \beta_1 x_{6P}$
Sixth Grade Advanced model: Y = $\beta_0 + \beta_1 x_{6A}$

Seventh Grade Needs Improvement model: Y = $\beta_0 + \beta_1 x_{7NI}$
Seventh Grade Proficient model: Y = $\beta_0 + \beta_1 x_{7P}$
Seventh Grade Advanced model: Y = $\beta_0 + \beta_1 x_{7A}$

Eighth Grade Needs Improvement model: Y = $\beta_0 + \beta_1 x_{8NI}$
Eighth Grade Proficient model: Y = $\beta_0 + \beta_1 x_{8P}$

Eighth Grade Advanced model: Y = $\beta_0 + \beta_1 x_{8A}$

In an effort to fully explore the relationship between teachers' use of the Frameworks to plan instruction and MCAS scores, the percent of students scoring in each of the Needs Improvement, Proficient, and Advanced performance categories were added together to create a new variable for each teacher. The percentages, as above, were those reported by DESE and aggregated across all teachers in a school by grade level. The combined scores in these performance categories all represent passing scores. A series of linear regressions was used to evaluate the relationship between teachers' scores on the Use Scale, the independent variable, and the new variable of combined percentages, the dependent variable, for each grade level. The regression equations for each model were:

Third Grade Combined: Y = $\beta_0 + \beta_1 x_{3Combo}$

Fourth Grade Combined: Y = $\beta_0 + \beta_1 x_{4Combo}$

Fifth Grade Combined: Y = $\beta_0 + \beta_1 x_{5Combo}$

Sixth Grade Combined: Y = $\beta_0 + \beta_1 x_{6Combo}$

Seventh Grade Combined: Y = $\beta_0 + \beta_1 x_{7Combo}$

Eighth Grade Combined: Y = $\beta_0 + \beta_1 x_{8Combo}$

CHAPTER 4

RESULTS

This chapter is organized in sections presenting the results of the pilot study and full study, respectively. Descriptive results are presented for the pilot study because the small sample size was not appropriate for statistical testing. The results of the full study address the specific research questions and are presented in the order the research questions are stated, preceded by the results of the factor analysis.

Pilot Study

The survey responses indicate there were 16 female and 7 male respondents, and one omitted gender response. The mean number of years of teaching experience was 12 years; however, responses ranged from 1 year to 28 years of teaching experience. All respondents taught at the High School (N = 20) or Middle School (N = 4) level. The majority of teachers responding to this survey, 92%, reported working in middle income (N = 13) to high income (N = 9), suburban school districts (N = 17). Two teachers reported working in a low income school district. Three teachers reported working in a rural school district, two teachers reported working in an urban school district, and two teachers omitted information about their school district.

There was a wide range of subjects taught among the teachers. Briefly summarized, there were five science teachers, four math teachers, and a mix of English teachers, Special Education teachers, arts teachers, and more. Please refer to Appendix A for a complete listing of the subjects taught by responding teachers.

This survey contained 31 questions. Reporting responses to each question would be time-consuming and, perhaps, tedious for the reader. Therefore, the results will be summarized. Please refer to Appendix A for a complete reporting of the results for every question on the survey, including verbatim responses to the open-ended questions.

The results of the questions related to attitude towards and use of the Curriculum Frameworks indicated that 92% of the teachers were Very Familiar (N = 9) or Familiar (N = 13) with the Curriculum Frameworks. Nineteen of the teachers (79%), when asked if they liked the Curriculum Frameworks, responded Yes. Three teachers, or 13%, reported that they do not like the Curriculum Frameworks and two teachers did not respond to the question (because there are no Curriculum Frameworks for their subject area).

The three teachers who answered No when asked if they like the Curriculum Frameworks were asked to explain their answers. Generally, their responses indicated that the content included in the Curriculum Frameworks was not what they considered most relevant (please refer to Appendix A for their verbatim responses).

Despite the favorable attitude toward the Curriculum Frameworks, only 5 teachers (21%) use them Very Often to plan classroom instruction, and 5 teachers (21%) reported using them Often. A total of 14 teachers (58%, including the two teachers teaching subjects without Curriculum Frameworks) reported using the Curriculum Frameworks Sometimes (N= 8) or Rarely (N = 6) when planning classroom instruction. Furthermore, only 25% of teachers reported that they follow the Curriculum Frameworks Very Closely when planning instruction. The remaining 75% of respondents either followed the Curriculum Frameworks Somewhat (N=7), used them as a Loose Guideline (N = 8), used them to Get Ideas (N = 2) or Did Not Use the Curriculum Frameworks to plan instruction (N = 1).

The results for this sample of Massachusetts public school teachers show that less than 50% of the teachers use the Frameworks to plan instruction often or very often. Along with that worrisome finding, only 25% of these teachers follow the Curriculum Frameworks very closely when planning instruction. These results suggest that there is cause for concern regarding the degree to which Massachusetts teachers use the Curriculum Frameworks to plan instruction. Further investigation is necessary to determine if these results are specific to this sample of Massachusetts teachers or if the results generalize to a broader sample of teachers.

Seventy-nine percent of teachers reported including topics/materials in instruction that are not in the Curriculum Frameworks Sometimes (N = 7), Often (N = 5), and Very Often (N = 8). The teachers who reported using topics/materials not in the Curriculum Frameworks were asked to explain why. The results broadly indicate, excluding the two teachers for whom no Curriculum Frameworks exist, that the teachers feel the need to include fun and relevant topics to keep the students interested. Furthermore, the responses seemed to indicate that the Curriculum Frameworks contain more breadth of content and the teachers see a need for more depth.

Finally, the teachers were asked if they ever feel overwhelmed by the Curriculum Frameworks. Over half of the teachers, 58%, answered positively; with 11 indicating Sometimes, 2 indicating Often, and 1 indicating Very Often. The teachers were asked what tools would be helpful in assisting them with using the Curriculum Frameworks to guide instruction. Professional development (54%), an alignment tool (46%), and an online tutorial from the MA DSE (25%) were the most popular responses. Yet, when asked if they would like help using the Frameworks to guide instruction, 71% of the teachers responded No.

Due to the small sample size and skewed distribution of the variables, significance testing was not conducted; however, there were some interesting patterns in the responses when the data was looked at according to gender and the level of experience of the teacher. One person did not respond to the question asking about gender, therefore his or her responses are not included in the following. Ninety-four percent of the female teachers indicated that they like the Curriculum Frameworks (one female teacher omitted a response). The six male teachers had mixed feelings about the Curriculum Frameworks. Three male teachers, or 50%, indicated they like the Frameworks and 50% did not like the frameworks (one male teacher omitted a response). Furthermore, the female teachers were much more likely to use the Curriculum Frameworks to plan instruction (Very Often = 4, Often = 3, Sometimes = 7, Rarely = 2) than were the male teachers (Very Often = 1, Often = 2, Sometimes = 0, Rarely = 4).

The range of years of teaching experience was made into a new variable, Level of Experience, and broken into three categories: New (1-5 years, N = 4), Experienced (6-15 years, N = 12), and Senior (16 years and higher, N = 8). A very interesting finding here is that the Experienced and Senior teachers were much *less* likely to use the Curriculum Frameworks to plan instruction, with over half of each group reporting using them only Sometimes or Rarely. This was not a result of being overwhelmed by the Frameworks, however, as again over half of the members in the Experienced and Senior groups Rarely or Never found the Curriculum Frameworks overwhelming. In contrast, all of the New teachers reported feeling overwhelmed by the Curriculum Frameworks Sometimes (N =

2), Often (N = 1), or Very Often (N = 1). This finding is borne out in the response to the question of would the teachers like help using the Frameworks. Three of the four New teachers responded yes while the majority of the Experienced and Senior teachers responded No. In fact, sixteen of nineteen Experienced and Senior teachers did not want help using the Curriculum Frameworks.

Full Study

Survey results will be reported only in relation to the research questions. A full summary of the results of the Teacher Survey and the Principal Survey can be found in Appendices F and G, respectively.

Many survey questions were used to measure teachers' attitude towards and opinions of the Frameworks, as well as the way teachers use the Frameworks to plan instruction. To allow for meaningful analyses, appropriate composite scores were created for each construct. An exploratory factor analysis (EFA) was used to construct subscales to be used to answer the research questions. The EFA revealed an initial solution containing three factors that accounted for 81.80% of the variance in the subset of survey questions. Unfortunately, the solution was difficult to interpret as the factor loadings of the questions on the third factor were very low, much lower in fact than on the first and second factors. A scree plot of the three factor solution indicates that, while there is a third factor present, it contributes minimally to understanding the factors underlying the questions. Please refer to Figure 3. This finding, coupled with the uninterpretable factor loadings, called for an analysis of a two factor solution.

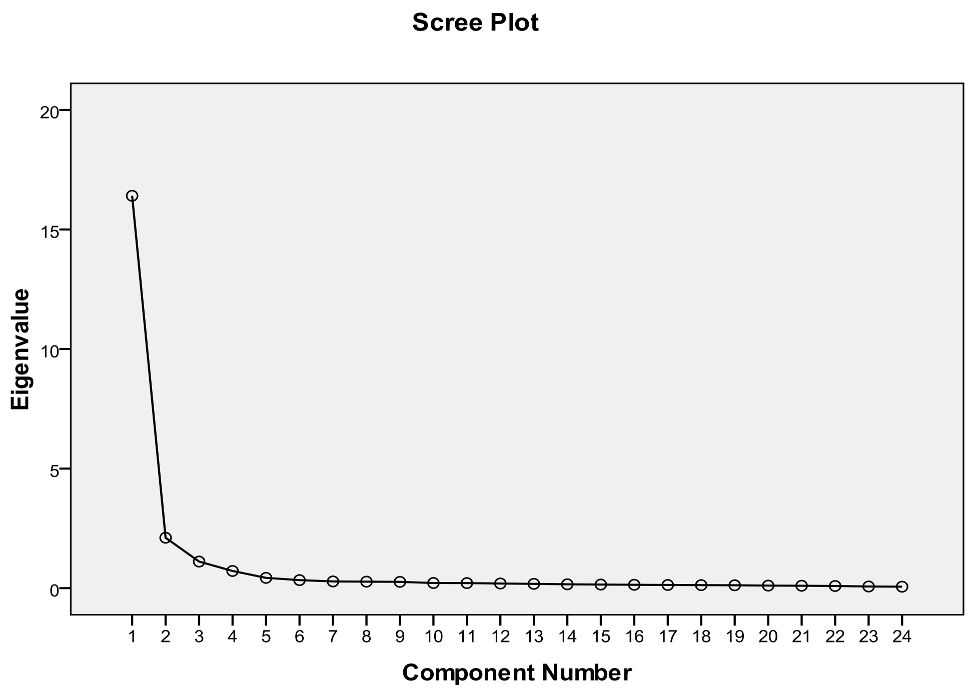

Figure 3. Scree plot for the three factor solution.

An EFA constrained to two factors explained 77.16% of the variance within the subscale questions and produced clearly defined factor loadings for each survey question, which allowed for a logical and reasonable interpretation of the results. Thus, the two factor solution was retained and subscales were constructed. Please refer to Appendix H for a list of survey questions and their factor loadings.

The two subscales that were created based on the EFA factor loadings, and the substantive interpretation of the subscales, were the Opinions Scale and Use Scale. The Opinion Scale is composed of the questions with high factor loadings on factor 1. These questions focus on teachers' opinions about the content and utility of the Math Frameworks as well their opinion about how much autonomy teachers should have when deciding what to teach. The questions included in the Use Scale focus on the degree to which teachers use the Frameworks to plan instruction and how the Frameworks are used.

An additional subscale was created upon further review of the survey questions. Although they fit together statistically, three questions on the Use scale were substantively different from the others in that they asked if the teachers engaged in practices that *support* the use of the Frameworks to plan instruction, rather than asking directly about their use of the Frameworks. Those three questions were separated out to form a third scale called Support. Table 3 presents descriptive information about the subscales.

Table 3. Subscale Information.

Subscale	Questions	Coefficient alpha	Interpretation
Opinions	6 - 21	0.47	Opinions about the Frameworks
Use	26, 27, 28, 30, 31	0.54	Use of the Frameworks
Support	29, 32, 33	0.62	Practices that support the use of the Frameworks

Unfortunately, as shown in Table 3 above, the Opinion Scale and Use Scale created based on the EFA failed to yield adequate reliabilities. Therefore, two subscales were created based solely on the substantive interpretation of the questions. The subscale themes remained opinions about the Frameworks and use of Frameworks. The Support Scale, which was a byproduct of the EFA, was not created again because it was not necessary for analysis of the research questions.

Once again, responses to the questions regarding opinions about and use of the Frameworks were combined to create one scale score per respondent. The Opinion Scale questions focus on teachers' opinions about the content and utility of the Math

Frameworks as well their opinion about how much autonomy teachers should have when deciding what to teach. The questions included in the Use Scale focus on the degree to which teachers use the Frameworks to plan instruction and how the Frameworks are used. The Support Scale was not considered

The reliability of each subscale was, once again, evaluated. The Opinion Scale produced a low reliability estimate and was deemed inappropriate for use in the analyses. Individual questions related to teachers' opinions of the Frameworks were determined to be more appropriate for use in the analyses. The reliability of the Use Scale is moderate; however, it was determined to be acceptable given the exploratory nature of this study. Therefore, the Use Scale was retained for use in analyses. Table 4 below presents descriptive information about the subscales.

Table 4. Revised Subscale Information.

Subscale	Questions	Coefficient alpha	Interpretation
Opinions	7-9, 12, 13, 16, 18-21	0.3	Opinions about the Frameworks
Use	6, 10, 11, 14, 15, 17, 26-28, 30, 31	0.6	Use of the Frameworks

Research question 1a. Are there variations in teacher practice related to the use of the Math Frameworks for planning and implementing classroom instruction? Descriptive statistics indicate there were differences in the degree to which teachers used the Math Frameworks to plan instruction. Additionally, there were substantial differences in teachers' opinions of the Math Frameworks and the degree to which professional development was used to support the use of the Math Frameworks. Table 5 below lists the range of values observed of the respondents on the Use Scale, as well as the mean and standard deviation of the scale. Please note this information includes responses from the 652 respondents with only one answer missing from survey responses. Respondents were removed from the data set used to calculate scale scores if more than one response was missing because the missing responses would have resulted in artificially low scale scores and misleading results.

Table 5. Table of Descriptive Statistics for Use Scale.

Scale	No. of Questions	Minimum	Maximum	Mean	Stand. Dev.
Use	11	11.00	50.00	36.77	4.74

The Use Scale contains scores across most of the scale range; the maximum possible scale score is 55 on the Use Scale. The score variability indicates that the degree to which teachers use the Frameworks to plan instruction varies.

The following histogram, entitled Figure 3, provides a graphical representation of the distribution of the scores for the Use Scale

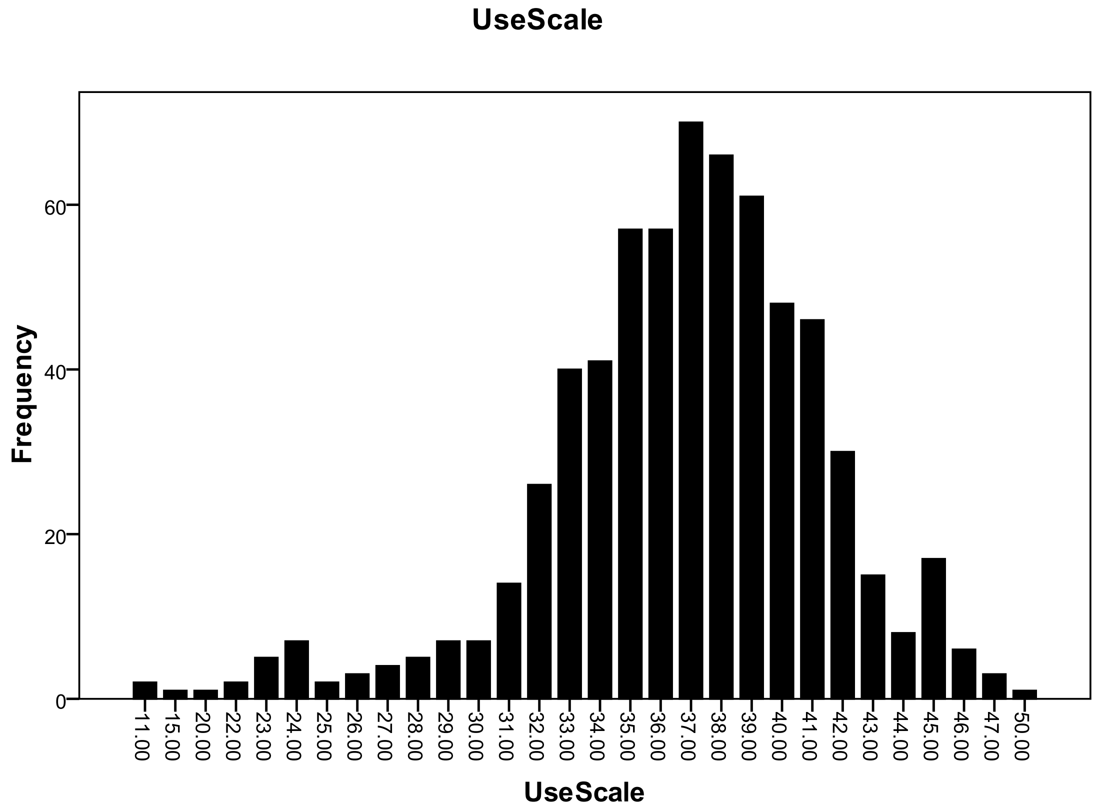

Figure 3. Score distribution of Use Scale.

Figure 3 shows a relatively normal distribution with a bit of negative skew. This is a positive finding that indicates the majority of the teachers were using the Frameworks to plan instruction. It also shows that the majority of the Use Scale scores are between approximately 33 and 42 points, indicating moderate to high use of the Frameworks to plan instruction.

The following Table, Table 6, provides response counts and the mean and standard deviation for the questions included in the Use Scale. This table provides information at the question level, rather than the scale level, which allows more attention to focus on the nuances of the responses. Please note the following table contains responses from all respondents regardless of their inclusion in the final scale.

Table 6. Use Scale Distribution, Means, and Standard Deviations.

Use Scale	Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree	Response Count	$\bar{x}$	sd
6. I am familiar with the Math Frameworks.	1%(8)	0%(3)	3%(18)	34%(236)	61%(420)	92%(685)	4.56	.679
10. I always use the Math Frameworks when developing a lesson plan.	3%(23)	17%(115)	24%(161)	36%(247)	20%(135)	91%(681)	3.51	1.111
11. I use the Math Frameworks mostly as a loose guideline when planning instruction.	9%(60)	28%(188)	23%(154)	37%(249)	5%(31)	92%(682)	3.00	1.080
14. There is not enough instructional time to cover the content of the Math Frameworks.	2%(14)	13%(86)	16%(107)	37%(248)	33%(222)	91%(677)	3.84	1.131
15. The Math Frameworks allow for in-depth coverage of content.	13%(92)	42%(287)	24%(162)	17%(115)	4%(24)	91%(680)	2.54	1.043
17. I follow the Math Frameworks closely when I plan instruction.	4%(25)	14%(97)	24%(161)	42%(284)	16%(110)	91%(677)	3.52	1.067

	Never	Rarely	Sometimes	Often	Always	Response Count	$\bar{x}$	sd
26. I use the Math Frameworks to plan instruction.	4% (28)	6% (37)	24% (158)	37% (241)	29% (188)	88% (652)	3.80	1.057
27. I refer back to the Math Frameworks for guidance during the course of planning instruction.	5% (33)	8% (55)	28% (179)	37% (241)	22% (142)	87% (650)	3.61	1.093

28. I include topics in instruction that are NOT included in the Math Frameworks.	6% (37)	23% (150)	50% (329)	17% (110)	4% (27)	88% (653)	2.91	.882
30. When planning instruction I create entire lessons with content that does NOT appear in the Math Frameworks.	36% (236)	44% (289)	15% (99)	3% (17)	2% (10)	87% (651)	1.88	.869
31. I provide my students with more information (theories, formulas, steps to follow, “tricks”, etc.) than is listed in the Math Frameworks.	4% (27)	7% (43)	30% (193)	44% (287)	16% (101)	87% (651)	3.60	.980

$\bar{x}$ = sample mean; sd = sample standard deviation

Sixty-eight percent of teachers reported using the Frameworks Often or Always to plan instruction. This is a surprisingly large proportion and indicates that there very likely are variations in the degree to which Massachusetts Math teachers use the Frameworks to plan instruction. The majority of teachers, 87%, also reported referring back to the Frameworks Sometimes (28%), Often (37%), or Always (22%) while planning instruction. Again, this indicates a high degree of use of the Frameworks but variations in the degree to which they are used.

Fifty percent of the teachers reported Sometimes including topics in instruction that are not included in the Frameworks, while 17% included additional topics Often and 4% included them Always. This may encompass the last question on the scale which addresses providing students with *more* information than is included in the Frameworks for a topic. Eighty-nine percent of teachers reported providing more information than is listed in the Frameworks Sometimes (30%), Often (44%), or Always (16%). Only 20% of teachers, though, reported planning entire lessons based on content that was not in the Frameworks Sometimes (15%), Often (3%), or Always (2%). This indicates that the majority of teachers use the Frameworks to plan instruction and sometimes provide supplemental information. These responses further support the research suggesting that the degree to which the Frameworks are used varies by teacher.

Table 7 below reports teacher responses to questions related to opinions about the Frameworks.

Table 7. Response distributions to questions related to opinions about the Frameworks, means, and standard deviations.

Survey Questions	Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree	Response Count	$\bar{x}$	sd
7. The Math Frameworks are helpful for instructional planning.	1%(6)	3%(18)	8%(55)	51%(350)	37%(257)	92%(686)	4.24	.752
8. The Math Frameworks are a good representation of the material that should be taught.	1%(4)	8%(53)	16%(110)	55%(373)	21%(143)	92%(683)	3.88	.853
9. All of the content included in Math Frameworks is equally important.	4%(27)	40%(273)	20%(136)	28%(191)	8%(58)	92%(685)	2.97	1.087
12. The content of the Math Frameworks appropriately represents what students need to learn.	2%(12)	17%(118)	23%(159)	50%(338)	8%(55)	92%(682)	3.46	.940
13. The Math Frameworks try to cover too much material.	2%(14)	16%(110)	22%(149)	34%(234)	25%(172)	91%(679)	3.64	1.125
16. The Math Frameworks are missing important content.	4%(25)	32%(215)	43%(288)	18%(124)	4%(24)	91%(676)	2.83	.924
18. The Math Frameworks are overwhelming.	5%(33)	27%(179)	33%(222)	27%(178)	9%(59)	90%(671)	3.02	1.113
19. I would like help with using the Math Frameworks to plan instruction.	14%(92)	38%(258)	26%(176)	18%(125)	4%(25)	91%(676)	2.58	1.071
20. I have no interest in using the Math Frameworks for instructional planning.	47%(316)	37%(250)	13%(90)	2%(15)	1%(8)	91%(679)	1.72	.852
21. Teachers should have complete autonomy when deciding what to teach.	42%(286)	41%(279)	12%(81)	4%(25)	2%(12)	92%(683)	1.81	.899

$\bar{x}$ = sample mean; sd = sample standard deviation

Table 7 provides responses to the questions related to teachers' opinions about the Frameworks. As you can see, opinions about the Frameworks vary; however, they were mostly favorable with 76% of the teachers believing the Frameworks are a good representation of the material that should be taught. Only 22% of the teachers responding to this survey believed the Frameworks are missing important content. There was more disagreement about the degree to which the Frameworks represents important material, with only 58% of teachers believing the Frameworks appropriately represent what students need to learn and only 36% of teachers agreeing that the content in the Frameworks is all equally important.

Roughly one third, 36%, of teachers reported feeling that the Frameworks were overwhelming. Furthermore, many of the teachers reported that they feel the Frameworks try to cover too much material, 59%, and they do not have enough instructional time to cover the contents of the Frameworks, 70%. However, only 22% of teachers wanted help with using the Frameworks to plan instruction.

It is interesting to note that only 3% of teachers were not interested in using the Frameworks to plan instruction and only 6% of teachers believed they should have complete autonomy when deciding what to teach. These results indicate that perhaps, for this group of teachers, low use of the Frameworks to plan instruction is not a result of personal opinions about autonomy.

These results indicate that the mostly positive opinions about the content of the Frameworks overall are tempered by conflicting opinions about the more nuanced contents of the Frameworks. Teachers appear to struggle with the breadth of the

Frameworks and this may impact the degree to which they plan instruction using the Frameworks.

Table 8 below reports teacher responses to questions related to the utility of teacher preparation programs, professional development, and programs designed to assist teachers use the Frameworks to plan instruction. It was surprising to see that 14% of teachers Never and 12% of teachers Rarely use a program designed to facilitate the use of the Frameworks to plan instruction. And only 20% of teachers Always use such a program. Given the importance placed on the use of the Frameworks, even prior to the proposal to use MCAS as indicators of teacher effectiveness, it is curious that programs designed to help with this are not more widely implemented. Similarly, information received from professional development is only used Often by 33% of the teachers and Always by 7% of the teachers. The majority of teachers use information received from professional development(s) designed specifically to assist teachers use the Frameworks to plan instruction only Sometimes (34%). Of course, only 68% of the respondents had actually participated in this type of professional development, so that should be kept in mind when thinking about these results. Finally, only 45% of teachers reported talking with their colleagues about ways to use the Frameworks to plan instruction Often (32%) or Always (13%). Again, a surprisingly low proportion given the emphasis on test scores.

Table 8. Responses to questions about teacher preparation, professional development, and programs designed to assist teachers with using the Frameworks to plan instruction, sample means, and standard deviations.

Survey Questions	Never	Rarely	Sometimes	Often	Always	Response Count	$\bar{x}$	sd
29. I use a program that facilitates the use of the Math Frameworks to plan instruction (e.g., curriculum mapping or some other program).	14% (88)	12% (81)	19% (121)	35% (229)	20% (132)	87% (651)	3.36	1.312
32. I use information received from professional development to plan instruction using the Math Frameworks.	11% (73)	14% (93)	34% (224)	33% (216)	7% (45)	87% (651)	3.10	1.101
33. My colleagues and I talk about ways to use the Math Frameworks in instructional planning.	9% (59)	18% (117)	28% (181)	32% (210)	13% (85)	88% (652)	3.22	1.160

$\bar{x}$ = sample mean; sd = sample standard deviation

Research question 1b. Are teacher behaviors related to the use of the Math Frameworks associated with gender and/or grade taught? The three assumptions associated with ANOVA are independence of observations, normal distribution of the dependent variable, and homogeneity of variance. These assumptions were evaluated to assess the appropriateness of using an ANOVA for this research question.

The assumption of independence of the observations is presumed to be met because individual teachers provided answers based on their own opinions and experience. The distribution of the Use Scale approximated a normal distribution. And, finally, a Levine's test for homogeneity of variance was not significant, $p = .217$, indicating the error variance of the Use Scale is equal across all groups. Failing to reject the null hypothesis is desirable here as it indicates an ANOVA is an appropriate statistical test to use on these data.

The dependent variable in the ANOVA was teachers' scores on the Use Scale. The independent variables were gender and grade taught. A new level of grade taught was created to accommodate teachers that taught multiple grade levels. All teachers teaching multiple grades, regardless of the grades, were assigned to this level. There were then seven possible levels for grade taught: grades 3 – 8 (six levels) and multiple grades.

A two-way ANOVA, with 2 x 7 or 14 levels, was run. The results showed the main effect for grade taught and the interaction of gender and grade taught were not significant; however, the main effect for gender was significant, $p < .05$.

The planned Tukey post hoc comparison was not necessary because the regression indicated there was no significant difference between the means of any of the grades taught.

The ANOVA table and the means and standard deviations for the dependent variable, the Use Scale, for each group are presented below.

Table 9. ANOVA Table for Research Question 1b.

Source	Degrees of freedom	F	Significance
Corrected Model	13	2.04	.016
Intercept	1	17550.616	.000
Gender	1	9.744	.002*
Grade	6	.332	.920
Gender * Grade	6	1.639	1.34
Error	631		
Total	645		
Corrected Total	644		

a. R Squared = .040 (Adjusted R Squared = .021)
* Significant at $p < .05$

Table 10. Means and Standard Deviations of Use Scale by Gender and Grade level.

Group	Mean	SD
Male	35.2128	5.22169
Female	37.0236	4.61080
All grades	36.7597	4.74370
Grade 3	36.9386	3.87934
Grade 4	37.0376	4.38387
Grade5	36.2742	4.65796
Grade 6	37.4638	5.47878
Grade 7	36.0208	4.90544
Grade 8	37.0000	5.26387
Multiple Grades	36.5556	5.27086

The results of the ANOVA, presented in the ANOVA Table in Table 9, indicated there was a significant difference in the degree to which the Frameworks were used by males and females, based on their scores on the Use Scale. Table 10 shows the means of each gender group. Based on these results, female teachers tend to use the Frameworks to plan instruction more often than male teachers. Although no significant differences were found between the mean Use Scale scores between teachers of different grades, the means and standard deviations are presented as well in Table 10.

The graph below, Figure 5, presents a plot of the estimated marginal means for male and female teachers' scores on the Use Scale by grade level. The estimated marginal means considers the mean of the dependent variable, Use Scale, across all levels of the independent variables, gender and grade taught, in relation to the sample size and without the associated error. This is important here because of the large difference between the number of male and female teachers. Parallel lines indicate no significant difference in the means across levels because they increase or decrease in the same way across groups. Although these lines are not parallel, the interaction was not statistically significant. The lines are not coincident, however, indicating that there is a significant difference in the Use Scale means for males and females across the levels of grades taught.

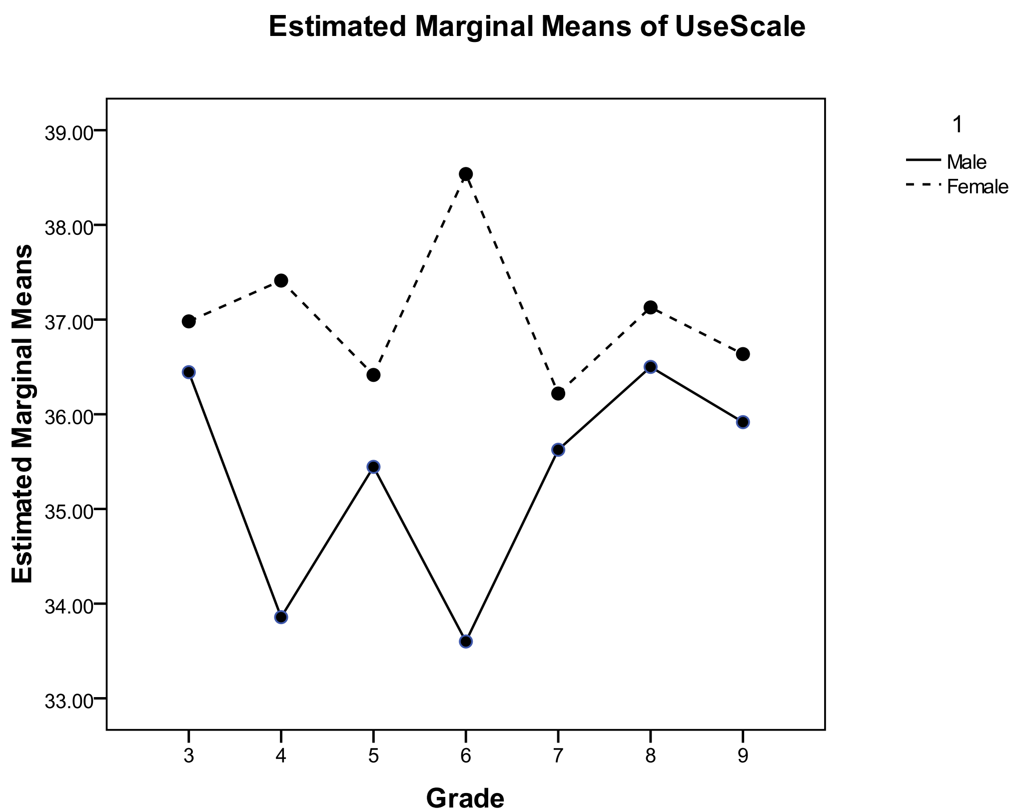

Figure 5. Estimated marginal means of Use Scale.

Given the significant differences in scores on the Use Scale based on gender, further analysis was conducted to determine if the differences were reflected in the MCAS scores of male and female teachers. Plainly stated, were students of female teachers more likely to receive higher MCAS scores than students of male teachers?

A chi-square test of independence with the variables gender and performance category, weighted by the proportion of students scoring in each performance category, was performed. The results were not significant and indicate no association between the gender of the teacher and the proportion of students scoring in each performance category, x^2(3, N = 591) = 1.087, p = 0.78.

Research questions 1c – 1g. Are years of teaching experience (1c.), teachers' opinions about autonomy (1d.), teacher preparation programs (1e.), participation in professional development (1f.), and opinions about the Frameworks (1g.) associated with use of the Frameworks? Correlations were used to explore the relationships among the variables addressed by each research question and the Use scale. Table 11 presents the results of the correlation study.

Table 11. Correlations Between Years Teaching, Autonomy, Teacher Preparation, Professional Development, Opinions About the Frameworks, and the Use Scale.

Research Question No.	Survey Question No.	Variable	Correlation with Use Scale*	95% Confidence Interval
1b	3. How long have you been teaching?	Years teaching	0.013	
1c	21. Teachers should have complete autonomy when deciding what to teach.	Autonomy	-0.013	
1d	22. My teacher preparation program adequately prepared me to plan instruction using the Math Frameworks.	Prep program	0.146*	0.071-0.22
1e	25. Were the Professional Development(s) you attended helpful in assisting you to use the Math Frameworks to plan instruction?	Prof. Dev.	-0.218*	-0.289 – -0.144
1f	7. The Math Frameworks are helpful for instructional planning.	Opinion	0.564*	0.51-0.614
1f	12. The contents of the Math Frameworks appropriately represents what students need to learn.	Opinion	0.184*	0.109-0.257
1f	18. The Math Frameworks are overwhelming.	Opinion	-0.017	

* Correlation significant at 0.01

The correlations between years of teaching experience, autonomy, and feeling overwhelmed by the Frameworks did not have a significant correlation with scores on the Use Scale, which indicate use of the Frameworks.

Scores on the Use Scale showed significant low, positive correlations with feeling that a teacher preparation program adequately prepared the teacher to use the Frameworks to plan instruction and opinions about the Frameworks appropriately representing what students need to learn.

There was also a significant low, negative correlation between scores on the Use Scale and feelings that professional development was helpful in assisting the teacher to use the Frameworks to plan instruction. The negative correlation indicates an inverse relationship between use of the Frameworks and feeling that professional development was helpful. This is a curious result; however, only 68% of respondents, or 463 teachers, actually attended a professional development designed to assist teachers with using the Frameworks to plan instruction. Of those teachers only 187 teachers found the professional development helpful. The negative correlation may be a result of the remaining 276 teachers reporting low opinions of the helpfulness of the professional development but still reporting high Use Scale scores. In this light, the results are not quite so curious because teachers report generally high Use Scale scores regardless of participation in professional development.

The observed significant correlations reported above do not indicate a particularly strong relationship between the adequacy of a teacher preparation program, attending professional development, or opinions about the appropriateness of the content of the Frameworks and the use of the Frameworks to plan instruction. In fact, feeling that their teacher preparation program adequately prepared them to plan instruction using the Frameworks explains only 2% of the variability in Use Scale scores. Attending a professional development that teachers felt was helpful in assisting them to plan instruction based on the frameworks explained 5% of the variability in Use Scale scores. Finally, feeling that the Frameworks appropriately represented what students need to learn explained 3% of the variability in Use Scale scores.

These findings may be a result of the subjectivity inherent in the questions. The correlation is actually between their feelings about the variable and their score on the Use Scale. Although low, the amount of variability in the Use Scale scores explained by these other variables must be considered. Therefore, we can say there is likely a slight relationship between teachers' feelings about their teacher preparation program, professional development(s), and their opinion about the appropriateness of the content of the Frameworks and their use of the Frameworks to plan instruction.

A stronger, though still moderate, correlation is observed between teachers' opinions about the Frameworks being helpful for instructional planning and their use of the Frameworks. The significant .564 correlation indicates that as teachers' opinions about the helpfulness of the Framework increase, their use of the Frameworks also increases, to a moderate degree. In fact, teachers' opinions about the helpfulness of the Frameworks for planning instruction account for 32% of the variability in Use Scale scores. This is a large amount of explained variability and indicates that teachers' opinions about the Frameworks may play a significant role in the degree to which they use the Frameworks to plan instruction.

Research question 2a.What is the relationship between the teachers' self-reported use of the Frameworks and student performance, as reported by the respondent? The relationship between teachers' use of the Math Frameworks and their students' scores on MCAS, as reported by teachers, was explored using a series of simple regressions with the teachers' score on the Use Scale as the independent variable and the percent of students scoring in the Needs Improvement, Proficient, and Advanced performance categories as the dependent variable. Results for each regression are reported in Table 12 below.

Table 12. Regression Table for Research Question 2a.

Model	Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	t	Sig.	95% Confidence Interval for B Lower	95% Confidence Interval for B Upper	Adjusted R^2**
			Needs Improvement				
Constant	15.789	6.303	2.505	.012	3.412	28.167	
Use Scale	.234	.170	1.377	.169	-.100	.567	
			Proficient				
Constant	51.988	6.957	7.472	.000	38.325	65.651	
Use Scale	-.011	.187	-.060	.952	-.379	.357	
			Advanced				
Constant	16.349	4.841	3.377	.001	6.843	25.856	
Use Scale	.044	.130	.337	.736	-.212	.300	

** Adjusted R^2 values are only reported for regression models found to be significant at $p < .05$.

None of the regression models were significant. These results indicate there is no relationship between the classification of students into each performance category based on MCAS scores, reported by the teacher, and the Use Scale score of the teacher.

Research question 2b. What is the relationship between a principals' attitude toward the Frameworks and MCAS scores for the school? As with Research Question

2a, a series of simple regressions was used to evaluate the relationship between a principals' opinion about the Math Frameworks and the percent of the students in the school scoring in the Needs Improvement, Proficient, and Advanced performance categories on MCAS, as reported by DESE. The independent variable here is question 9 from the Principal's Survey: The Math Frameworks are a good representation of the material that should be taught. Results for each regression are reported in Table 13 below.

Table 13. Regression Table for Research Question 2b.

Model	Unstandardized Coefficients B	Std. Error	t	Sig.	95% Confidence Interval for B Lower	Upper	Adjusted R^2**
			Needs Improvement				
Constant	32.977	5.733	5.752	.000	21.617	44.337	
Use Scale	-1.318	1.397	-.943	.348	-4.085	1.450	
			Proficient				
Constant	32.998	5.287	6.241	.000	22.521	43.475	
Use Scale	1.084	1.288	.841	.402	-1.469	3.636	
			Advanced				
Constant	16.235	8.195	1.981	.050	-.004	32.475	
Use Scale	1.841	1.996	.922	.358	-2.115	5.798	

** Adjusted R^2 values are only reported for regression models found to be significant at $p < .05$.

None of the regressions are significant. There does not appear to be a statistically significant relationship between the MCAS Math scores of students in a school and the opinion of the principal of the school about the Math Frameworks.

Research question 2c. What is the relationship between teachers' use of the Frameworks and MCAS Math scores, by grade level? This research question also uses a series of simple regressions to evaluate the percent of students scoring in the Needs Improvement, Proficient, and Advanced performance categories, reported by DESE, aggregated across all teachers in a school at each grade level.

To present a clearer picture of the relationship between scores on the Use Scale and the percent of students scoring in each performance category, the means and standard deviation for each performance category are presented in Table 14 and the correlation between the Use Scale and each performance category by grade level is reported in Table 15.

Table 14 shows that there tends to be a higher mean number of students classified in the Proficient performance category. There is also a large jump in the mean number of students classified as Advanced from grades 3 and 4 to grades 5 – 8. This may be a result of students becoming familiar with the MCAS over repeated administrations, an increase in student awareness of the importance of MCAS scores, or a difference in the way in which teachers and administrators prepare students to take the MCAS from earlier grades to the middle grades. Regardless of the cause, it may prove difficult to use MCAS scores as indicators of teacher effectiveness if the majority of students in the population are classified as Proficient or Advanced yet there are wide variations in teacher practice related to the use of the Frameworks to plan instruction (as indicted by the variability in Use Scale scores). The standard deviations appear comparable except for the Advanced performance category, where the larger standard deviations indicate more variability in the number of students classified as Advanced. Perhaps this area, at the higher end of the score scale, represents teachers who use of the Frameworks to plan instruction more often or an area where the degree of use of the Frameworks to plan instruction is more readily apparent in student scores. If such a relationship can be demonstrated, it may indicate MCAS scores should be considered as an indicator of teacher effectiveness.

Table 14. Means and Standard Deviations of Performance Categories by Grade Level.

	Performance Category					
	Needs Improvement		Proficient		Advanced	
Grade	Mean	SD	Mean	SD	Mean	SD
3	21.33	9.61	54.95	10.70	15.76	9.42
4	38.72	13.48	34.90	10.36	18.02	12.38
5	23.52	9.37	35.57	7.02	29.07	15.79
6	23.35	7.90	34.30	6.15	30.02	12.53
7	27.38	8.42	33.88	8.13	20.38	12.16
8	24.69	8.87	31.58	6.10	28.67	15.04

Table 15 indicates that the majority of the correlations between teachers' scores on the Use Scale and the percent of students classified in each performance category are not significant. In fact, only two of 18 correlations are significant. These two significant correlations are less than .3, which does not indicate a strong relationship between scores on the Use Scale and the percent of students classified in the performance category. However, scores on the Use Scale explain 6% and 8% of the variability in the percent of students scoring in the sixth grade Needs Improvement and Advanced performance categories, respectively. There appears to be a relationship, at least for these two sixth grade performance categories, which may be an indication that MCAS scores can be used as indicators of teacher effectiveness.

Table 15. Correlations Between Use Scale and Percent of Students in Each Performance Category.

Grade	Needs Improvement	Proficient	Advanced
3	0.04	-0.02	0.08
4	0.07	-0.11	-0.14
5	-0.02	0.10	-0.10

6	0.24*	-0.10	-0.29*
7	0.24	-0.08	-0.23
8	0.12	0.03	-0.16

*Significant at p < .05

Simple linear regression analyses of the relationship between the Use Scale and the percent of students scoring within each performance category by grade level produced only one significant result, out of 18 models, for the sixth grade Advanced performance category. The adjusted R^2 value indicates that 7% of the variability in the percent of students scoring in that performance is explained by teachers' scores on the Use Scale. Please refer to Table 16 below for a Regression Table that includes all of the linear regressions.

Table 16. Regression Table for Research Question 2c.

Model	Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B		Adjusted R^2**
	B	Std. Error			Lower	Upper	
Third Grade – Needs Improvement							
Constant	21.196	1.043	20.317	.000	19.125	23.266	
Use Scale	.002	.005	.359	.720	-.008	.012	
Third Grade – Proficient							
Constant	55.021	1.161	47.402	.000	52.717	57.324	
Use Scale	.000	.006	-.167	.867	-.012	.010	
Third Grade - Advanced							
Constant	15.478	1.020	15.170	.000	13.453	17.503	
Use Scale	.004	.005	.747	.457	-.006	.013	
Fourth Grade – Needs Improvement							
Constant	30.392	10.967	2.771	.007	8.653	52.131	
Use Scale	.223	.291	.765	.446	-.355	.800	
Fourth Grade – Proficient							
Constant	44.667	8.403	5.316	.000	28.011	61.323	
Use Scale	-.261	.223	-1.170	.244	-.703	.181	
Fourth Grade – Advanced							
Constant	32.279	10.011	3.224	.002	12.436	52.122	

Use Scale	-.381	.266	-1.434	.154	-.908	.146	
Fifth Grade – Needs Improvement							
Constant	23.572	.990	23.802	.000	21.608	25.537	
Use Scale	.000	.006	-.148	.882	-.012	.010	
Fifth Grade – Proficient							
Constant	35.295	.738	47.812	.000	33.831	36.759	
Use Scale	.004	.004	1.017	.312	-.004	.012	
Fifth Grade – Advanced							
Constant	29.682	1.661	17.872	.000	26.387	32.976	
Use Scale	-.009	.009	-1.020	.310	-.028	.009	
Sixth Grade – Needs Improvement							
Constant	5.870	9.419	.623	.536	-12.984	24.725	
Use Scale	.456	.244	1.866	.067	-.033	.944	
Sixth Grade – Proficient							
Constant	40.120	7.505	5.346	.000	25.097	55.143	
Use Scale	-.152	.195	-.780	.439	-.541	.238	
Sixth Grade – Advanced							
Constant	63.747	14.710	4.334	.000	34.302	93.191	**0.07***
Use Scale	-.879	.381	-2.306	.025	-1.642	-.116	
Seventh Grade – Needs Improvement							
Constant	8.082	13.037	.620	.539	-18.309	34.474	
Use Scale	.521	.351	1.487	.145	-.188	1.231	
Seventh Grade – Proficient							
Constant	40.020	12.906	3.101	.004	13.893	66.147	
Use Scale	-.166	.347	-.479	.635	-.869	.536	
Seventh Grade – Advanced							
Constant	47.464	18.853	2.518	.016	9.298	85.631	
Use Scale	-.732	.507	-1.444	.157	-1.758	.294	
Eighth Grade – Needs Improvement							
Constant	17.240	9.230	1.868	.069	-1.374	35.855	
Use Scale	.203	.249	.815	.419	-.299	.705	
Eighth Grade – Proficient							
Constant	30.503	6.392	4.772	.000	17.612	43.394	
Use Scale	.029	.172	.170	.866	-.318	.377	
Eighth Grade – Advanced							
Constant	44.493	15.575	2.857	.007	13.082	75.903	
Use Scale	-.431	.420	-1.027	.310	-1.278	.416	

** Adjusted R^2 values are only reported for regression models found to be significant at $p < .05$.

Looking at the relationship between teachers' scores on the Use Scale and the percent of students classified in each performance category, seventeen of eighteen regression models were not significant. This indicates that, for the 17 models, the teachers' scores on the Use Scale did not explain any of the variability in the scores for each performance category to a statistically significant degree.

The only significant regression model was flagged for sixth grade scores in the Advanced performance category. This regression model indicates a relationship between the Use of the Frameworks and MCAS grades and variations in the degree to which the Frameworks are used (represented by scores on the Use Scale) explain 7% of the variation in sixth grade, Advanced MCAS scores. Although the model only explains 7% of the variability in the Advanced scores, which seems a relatively small amount, it is not trivial. The fact that the model was able to account for 7% of the variability in MCAS scores takes on even more significance when considering how many factors are known to influence student test scores. This finding demonstrates a significant relationship between teachers' use of the Frameworks to plan instruction and students' scores in the model. This finding also suggests that , if MCAS scores are influenced by the teachers' use of the Frameworks to plan instruction, perhaps MCAS scores are an appropriate indicator of teacher effectiveness.

In an effort to fully explore the relationship between MCAS scores and use of the Math Frameworks, the percent of students scoring in each of the Needs Improvement, Proficient, and Advanced performance categories were added together to create a new variable for each teacher. The scores in these performance categories all represent passing scores, so the relationship between the Use Scale and the percent of students achieving passing MCAS scores was evaluated.

Table 17 reports the results of a series of simple regressions on the new variable of combined scores, which encompasses all passing MCAS scores, below. There are two significant relationships between the Use Scale and MCAS scores, which are indicated by the value listed in the Adjusted R^2 column.

Table 17. Regression Table for Recalculated Variable Combining MCAS Performance Categories.

Model	Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B		Adjusted R^2*
	B	Std. Error			Lower	Upper	
Third Grade							
Constant	91.695	1.124	81.589	.000	89.464	93.925	
Use Scale	.004	.005	.839	.404	-.006	.015	
Fourth Grade							
Constant	107.338	7.018	15.294	.000	93.426	121.249	**0.04***
Use Scale	-.420	.186	-2.253	.026	-.789	-.050	
Fifth Grade							
Constant	88.549	1.036	85.439	.000	86.493	90.605	
Use Scale	-.006	.006	-1.052	.295	-.018	.005	
Sixth Grade							
Constant	109.737	9.424	11.645	.000	90.873	128.601	**0.02***
Use Scale	-.575	.244	-2.355	.022	-1.064	-.086	
Seventh Grade							
Constant	95.566	20.241	4.721	.000	54.591	136.542	
Use Scale	-.377	.544	-.692	.493	-1.479	.725	
Eighth Grade							
Constant	92.236	12.001	7.685	.000	68.033	116.439	
Use Scale	-.199	.324	-.615	.542	-.851	.454	

* Adjusted R^2 values are only reported for regression models found to be significant at $p < .05$.

The models for fourth and sixth grades were each significant. The fourth grade model has an adjusted R^2 of .04 and the sixth grade model has an adjusted R^2 of .02. This result indicates that 4% and 2% of the variability in the fourth and sixth grade MCAS scores can be attributed to the teachers' scores on the Use Scale for this model, respectively. Unfortunately, this is not consistent across all of the regression models using the combined score variable. The regression models for all of the other grades were not significant, indicating no statistically significant relationship between Use Scale scores and the percent of students with passing MCAS scores. The two significant regression models should not be dismissed, however, because the nature of the data (survey, self-report) used in the analyses is not ideal, thus the significant findings may take on even more importance in determining if MCAS scores are valid indicators of teacher effectiveness.

CHAPTER 5

DISCUSSION

There are two overarching research questions guiding this study: 1.) do Massachusetts teachers teach the content of the Frameworks and 2.) if variations in teacher practice related to the use of the Frameworks exist, is the MCAS sensitive to these variations such that they are reflected in student performance? These questions will be evaluated in light of the survey results.

It appears, based on these results and those of the pilot survey data, that there is reason to believe there are variations in teacher practice relative to the degree to which the Math Frameworks are used to plan instruction. The range of scores observed on the Use Scale range from 11 to 50, which covers almost the entire width of the scale. This is not a particularly surprising finding, given the previous research on the topic and that teacher practice is bound to vary as do teachers' individual personalities and experiences (AFT, 2009; Porter, 2002; Reality Check, 2000; Remillard & Bryans, 2004; Sherin & Drake, 2010).

It is interesting to note, though, that Massachusetts teachers appear to use the Frameworks to a greater degree than the teachers included in previous research. That is to say, in the Reality Check Survey (2000) only 42% of responding teachers received most of their guidance about what they should teach from content standards. In Massachusetts, 56% of teachers reported that they always use the Frameworks when developing a lesson plan. This may be a result of time passing, as the reality Check Survey was conducted at the very beginning of standards- based reform efforts. Over the more than decade that has passed, the field of education has had a chance to adjust practice, teacher preparation programs, and introduce professional development.

Furthermore, this finding may be a result of changes over time in feelings related to teacher autonomy. Hamilton (2008) cites feelings of autonomy as a large factor in teachers' use of the Frameworks. The results of this survey show that only 3% of responding teachers feel teachers should have complete autonomy when deciding what to teach. This may also signal a shift in attitude that has occurred over time. Regardless of the cause, the Massachusetts teachers responding to this survey have more positive opinions about the Frameworks (76% positive overall) and do not appear to feel entitled to the same degree of autonomy as teachers participating in past research. Either, or both, of these factors may be contributing to the high degree of teachers' use of the Frameworks to plan instruction seen in this sample of responding teachers. Future research should focus on a more in-depth understanding of teachers' opinions about the content standards they are expected to use and their feelings about autonomy in the classroom in an attempt to better understand how to create an environment in which it is more likely teachers will use the content standards as they are intended to be used.

Of more interest is the gender effect discovered here. The pilot survey results hinted that there were differences in teacher practice related to gender; however, the sample size was too small to perform statistical testing. The full study indicates statistically significant differences in the degree to which male and female teachers use the Frameworks to plan instruction, with females reporting higher use than males.

This topic bears further investigation. Practically speaking, what does this mean for teacher effectiveness? If effective teachers plan their instruction using the Frameworks and male teachers use the Frameworks less than female teachers, are they less effective teachers? Are MCAS scores lower for students with male teachers?

Intuitively, this line of reasoning seems faulty. Statistically, a chi-square test of independence found no association between the gender of a teacher and the proportion of students scoring in each performance category.

The degree to which male and female teachers use the Frameworks to plan instruction and MCAS scores for male and female teachers should be explored for differences. This avenue of inquiry should be pursued because if male teachers actually do use the Frameworks to plan instruction less than female teachers but there is no difference in student test scores based on the gender of the teacher, it raises questions about the sensitivity of MCAS scores to variations in teacher practice, thus questions about the appropriateness of using MCAS scores as an indicator of teacher effectiveness.

There was a large difference in the number of men and women responding to this survey. Future research on this topic should prioritize obtaining information from larger samples of male and female teachers that reflect the proportions of each in the population of teachers. Unfortunately, Massachusetts either does not track or does not report the ratio of male to female teachers in the state; however, there is a national trend in public education whereby there is a larger, often much larger, proportion of female teachers. The staffing (not teacher level) information reported by the State shows there are 98, 523 female staff members and 24, 355 male staff members in the DESE. This suggests that Massachusetts likely follows the nationwide trend.

There is also a subtle but very interesting finding in the questions asked to explore variations in using the Frameworks to plan instruction. Looking at the questions in the Use Scale, 89% of teachers reported including topics in instruction that are not in the Frameworks. The teachers clearly feel these are important topics and that the information will help the students understand and master the content. But these topics are not in the Frameworks. What are the implications of this practice on effective teaching if effective teaching is defined as teaching the Frameworks? If the additional information increases student learning, are they still less effective teachers because they deviated from the Frameworks? Or are they only less effective if their students' MCAS scores are lower? And does the State really intend to prevent teachers, their content experts, from providing information to students they believe will enhance student learning?

The current study contradicts the results of the pilot study that years of teaching experience and opinions about autonomy affect teachers' use of the Frameworks to plan instruction. Neither of these factors were found to be significantly related to the use of the Frameworks to plan instruction. This is not altogether surprising given that this is a sample of math teachers for grades in which MCAS is given. It would simply be unfair to their students if they did not consider the Math Frameworks when planning instruction.

Teacher preparation programs, professional development, and teachers' opinions about the Frameworks were all significantly correlated to the use of the Frameworks. The correlations related to teacher preparation and professional development were low, which was initially a surprise. It seems that the quality of teacher preparation programs and professional development should explain more than 2% and 3% of the variability in Use Scale scores. Upon further reflection, however, it is possible that the same factors are at work here as with years of teaching experience and feelings about autonomy, which did not have a significant correlation with use of the Frameworks to plan instruction. Regardless of the years of teaching experience, feelings about autonomy, the effectiveness of a teacher preparation program, or professional development, teachers are likely using the Frameworks to the best of their ability simply because it benefits their students, at least in relation to MCAS.

It was less surprising to find a moderate correlation between teachers' opinions about the Frameworks and their use of the Frameworks. When teachers felt that the Frameworks appropriately represent what students need to learn and when they feel the Frameworks are helpful for instructional planning, their use of the Frameworks to plan instruction increases. In fact, the variability in teacher's opinions about the Frameworks explained 32% of the variability in Use Scale scores. This is a very large amount of explained variability and seems indicative of an important relationship between a teacher's opinion of the Frameworks and the degree to which a teacher uses the Frameworks to plan instruction. Again, this is not surprising but it has implications for the content of the Frameworks. If teachers do not feel the content of the Frameworks is appropriate, perhaps it is an indication that the State should reevaluate the content. Teachers are, after all, experts on the subject. This point becomes moot in 2013 when the Common Core State Standards will be fully adopted and implemented in Massachusetts; however, it would be interesting to know if teachers' opinions about the Common Core standards change the degree to which they use the standards to plan instruction.

Correlation never proves causation, so it is difficult to interpret the effect of teacher preparation programs, professional development, and opinions about the Frameworks on a teachers' tendency to use the Frameworks based on these low to moderate correlations. It is clear, though, that a relationship exists between these factors and the degree to which teachers use the Frameworks to plan instruction.

The correlations reported above, which indicate a relationship likely exists between teacher preparation programs and professional development and the degree to which teachers use the Frameworks to plan instruction raise concerns about the way teachers are prepared for the profession. Only 39% of teachers felt their teacher preparation program adequately prepared them to use the Frameworks to plan instruction. Only 68% of teachers received professional development designed to help them use the Frameworks to plan instruction; these findings are in line with previous research into these areas (AFT, 2009). There is a disconnect here between the task teachers are asked to perform and the tools they are provided with to perform the task. Previous research has brought this disconnect to light and discussed the impact on students. Now, by attaching consequences such as pay and employment to the same task, the disconnect is magnified. It is quite simply, to use the words you might hear a young student use, not fair.

When the State intends to define teacher effectiveness as the ability of the teacher to teach the Frameworks, regardless of how that is measured, it has an obligation to provide teachers with the tools to be effective. Otherwise teachers are being set up to fail. Future research, and the State of Massachusetts, must focus on ways to ensure that using content standards to plan instruction becomes a pivotal (and effective) part of the curriculum in teacher preparation programs. Professional development in this area is also a must for the thousands of teachers that entered the classroom before content-based educational reform came into effect. The benefit will be dual pronged because in helping our teachers we help our students. And, in the field of education, ultimately it is the students we are here to serve.

The second research question pertained to the ability of MCAS to reflect variations in teachers' use of the Frameworks, if such variations exist. The results of the first research question indicate it is likely such variations do exist. To explore the relationship between MCAS scores and variations in teacher practice, a series of simple linear regressions was used. The data used as a proxy for actual MCAS scores came from teacher self report and DESE report of the percent of students scoring in each of the three performance categories considered passing. The results of all the tests were very similar. The majority of the statistical analyses revealed a non-significant relationship between teachers' score on the Use Scale and their student scores on the MCAS Math test. There were only three, of twenty-seven, significant results; however, in each case the amount of variability in the MCAS scores that was explained by the model was relatively low. Two of the three significant tests indicated less than 2% and 4% of the observed variability in scores could be explained by the scores on the Use Scale. The third test indicated 7% of the variability in MCAS scores was explained by the scores on the Use Scale. This does not explain a great deal of the variability in scores but it certainly indicates a relationship exists. The relationship may prove even stronger, perhaps with more significant statistical tests or larger amounts of score variability explained by teachers' use of the Frameworks to plan instruction, if the appropriate data is obtained.

The fact that there was a significant relationship that explained some of the variability in students scoring in different performance categories is worth further consideration, particularly given the nature of the data used in these analyses; however, they do not explain enough of the variation in MCAS scores to support their use as indicators of teacher effectiveness at this point. When considered along with the many non-significant results, even the models able to explain some variability in the scores based on teachers' use of the Frameworks, these results indicate that student MCAS scores may not be an appropriate indicator of teacher effectiveness.

These results indicate that many more factors than can be accounted for simply by teachers' use of the Frameworks to plan instruction influence student MCAS scores. This is no surprise as the literature is rife with research indicating the many student-related factors (diet, educational level of parents, etc.) and teacher-related factors (years of experience, educational background, resources provided by the school district, etc.) influence student test scores. This is, if you recall, the purpose of VAMs. None of these factors were considered within this study because, unfortunately, the State's current plan for using SGPs to evaluate teacher effectiveness does not consider the influence of those factors on teacher effectiveness either. SGPs are calculated based solely on MCAS scores from one year to the next and school level demographic data. As the State has not defined which school level demographics will be considered, it was impossible to include questions related to this on the survey. Inclusion of those demographic variables may have led to a more robust regression model that explained a larger proportion of the variability in student MCAS scores.

Even if the demographic variables were known here, or if a valid VAM can be constructed, there is still a practical question that must be answered: How much variance in student scores related to or explained by the teacher is enough to judge effectiveness? Based on the results found here, can we judge an effective teacher based on 2%, 4%, or even 7% of the variance explained by teachers' use of the Frameworks to plan instruction (the de facto State definition of an effective teacher)? Is 50% of the variability in scores attributed to the teacher enough to make a judgment? 93%? There is no scale for this. There is also no research to support this practice. Therefore, the State should exercise extreme caution when using MCAS scores to make high-stakes decisions that affect our teachers.

This study in no way provides conclusive proof that MCAS scores do *not* reflect some degree of teacher effectiveness; however, the results are a caution that the relationship between student test scores and the effectiveness of the student's teachers is complicated and likely influenced by many factors. These results are supported by previous research on student test scores and current research into the utility of student growth models.

The lack of a relationship between a teacher's use of the Frameworks to plan instruction and student scores on MCAS found in this study may well be a result of an attempt to use a blunt instrument, a survey, to collect information about a topic with fine grains of variation. Future research should focus on ways to identify and accurately capture the information that will provide answers to these questions. The SEC (Porter & Smithson, 2001) is an existing tool that drills down to the level of detail necessary to properly evaluate degrees of variation in teacher practice. Another tool could be used or developed. Such a study would need to link the results from the participating teachers to their students' test scores to fully explore the relationship DESE posits. Although it has been called expensive and time consuming, an alignment study such as the SEC is necessary for the State to prove its assertion that student test scores are an appropriate measure of teacher effectiveness and defend the consequences, positive and negative, for teachers.

There were two topics of interest in the survey that, although not directly related to the research questions, deserve attention. These topics are related to the use of MCAS results and opinions about effective methods of teacher evaluations. Please see Table 18 below, which contains the questions related MCAS results and potentially useful methods of teacher evaluation that were common to both the Teacher and Principal Surveys. Response distributions from teachers and principals are also included in Table 18.

Table 18. Common Survey Questions and Response Distributions.

MCAS Results						
	Never	**Rarely**	**Sometimes**	**Often**	**Always**	**Response Count**
I spend a lot of time looking at MCAS Math results soon after I receive them.						
Teachers	2% (16)	9% (60)	20% (127)	31% (202)	38% (246)	87% (651)
Principals	0% (0)	2% (2)	9% (11)	38% (49)	52% (66)	82% (128)
After I get MCAS Math results for the school, I evaluate student performance with respect to the content strands in the Math Frameworks.						
Teachers	6% (36)	10% (68)	22% (142)	29% (190)	33% (214)	87% (650)
Principals	2% (2)	2% (2)	8% (10)	40% (51)	49% (62)	81% (127)

Methods of Teacher Evaluation					
	Not at all useful	**Minimally useful**	**Useful**	**Very useful**	**Response Count**
Peer observation					
Teachers	6% (35)	17% (108)	51% (321)	26% (167)	85% (631)
Principals	0% (0)	13% (17)	55% (69)	32% (40)	81% (126)
Principal observation					
Teachers	4% (25)	24% (151)	57% (361)	15% (97)	85% (634)
Principals	0% (0)	4% (5)	53% (67)	43% (54)	81% (126)
Outside evaluator					
Teachers	13%	38%	41%	8%	85%

	(81)	(240)	(259)	(52)	(632)
Principals	4% (5)	33% (41)	54% (67)	10% (12)	80% (125)
Test scores					
Teachers	16% (104)	47% (300)	32% (205)	4% (23)	85% (632)
Principals	2% (3)	27% (34)	61% (76)	10% (12)	80% (125)
Parent evaluations					
Teachers	31% (195)	50% (314)	19% (120)	1% (5)	85% (634)
Principals	13% (16)	62% (78)	24% (30)	2% (2)	81% (126)
Student evaluations					
Teachers	17% (108)	41% (256)	35% (223)	7% (45)	85% (632)
Principals	4% (5)	38% (48)	48% (60)	10% (12)	80% (125)
Videotaping the teacher during instruction					
Teachers	13% (80)	29% (182)	43% (273)	15% (94)	84% (629)
Principals	1% (1)	13% (16)	46% (57)	41% (51)	80% (125)

Teachers and Principals were asked about the ways they may use student MCAS results. It was interesting to see that 90% of principals reported spending a lot of time looking over MCAS results but only 69% of teachers spend a lot of time with the results. Similarly, 89% of principals evaluate student performance with respect to the Frameworks but only 62% of teachers report the same. This may reflect a clear separation of priorities between the two groups. Principals must be concerned with AYP reporting and how this may affect their budget and even their ability to run their school. Teachers, meanwhile, have faced minimal consequences due to MCAS scores up to this point.

Teachers have not been ignoring MCAS results. In fact, only 4% report ignoring results and 75% report being eager to receive their students' MCAS results; however, the degree to which the remaining 96% of teachers use the results is unclear and deserves further study. It is likely expected that teachers will use the results to guide future instruction, meaning that weak areas of performance receive more attention in future instruction. Yet only 62% of teachers evaluate student performance relative to the Frameworks. This finding raises questions that adjustments to future instruction occur with the regularity the State intended. It seems likely that teachers will be more in touch with MCAS results once they are used as an indicator of teacher effectiveness since they will be impacted directly.

Unfortunately, we do not know if they possess the tools needed to use the scores in a way that promotes student achievement. How exactly are teachers supposed to look at a list of items related to a specific standard and the number of students correctly answering the item, and develop an effective change in instruction? They will have to make decisions such as: there are 2 items aligned to this standard and most students got the items wrong – do I spend more time teaching the Standard? It seems that, once again, teachers will be faced with a task they have not been prepared to perform.

The second topic of interest is teacher and principal opinions about different methods of teacher evaluation. This topic is, in my opinion, crucial to examine because teachers and principals are in the best position to observe the success or failure of the proposed methods. There was some consensus among the teachers and principals about the best methods of teacher evaluation but there were also some differences.

Both teachers and principals agreed that peer observation or principal observation were the most useful methods of teacher evaluation; however, 77% of teachers reported that peer observation was the most useful method of evaluation and only 72% of teachers felt principal observation was useful. In contrast, 87% of principals felt peer observation was useful while 96% of principals felt principal evaluation was the most useful method of teacher evaluation. An outside evaluator was thought to be useful by 49% of the teachers and 64% of the principals. These results suggest that both groups feel that observation of operational teacher practice is the most useful form of teacher evaluation, but they disagree as to who is the best observer.

The disparity in the choice of observer is reflected in responses related to videotaping the teacher during instruction. Many more principals felt this was a useful method of evaluation (87%) than teachers (58%). The person(s) reviewing the videotape was not specified in the survey question. The responses were likely influenced by whomever the respondents assumed would review the videotape. Teachers assuming a principal or outside evaluator would review the tape would be less receptive to the method based on their responses above, Principals, on the other hand, may assume they themselves will review the videotape; this would explain the lower teacher ratings and higher principal ratings. Regardless of the reasons, videotaping instruction is not felt to be the most useful method of teacher evaluation in either group.

Parent and student evaluations were less warmly received. Only 20% of teachers and 26% of principals felt parent evaluations were a useful method of evaluating teachers. This is not surprising given that parents are removed from the classroom and have little other than students' accounts of the teacher and student grades to base evaluations upon. Student evaluations were thought to be more useful by both teachers and principals. 42% of teachers and 58% of principals reported student evaluations are a useful method of teacher evaluation.

Results related to the role of test scores in teacher evaluation were very interesting. Only 36% of teachers felt test scores were a useful method of evaluating teachers, yet 71% of principals felt they were useful. The number of respondents who feel test scores are a useful method of evaluating teachers is almost *double* for the principals. The very large disparity in the results illustrates that this method of teacher evaluation results in the most disagreement among teachers and principals. It would be very informative to better understand why the disparity exists. Many principals are former classroom teachers and, one might expect, would be sympathetic to the plight of the teachers if they were being subjected to an unfair method of evaluation. Furthermore, if principals thought using MCAS scores as an indicator of teacher effectiveness was inappropriate, they run the risk of losing good teachers. This could impact their entire school because the school is judged on MCAS results. Why then do principals appear to favor this practice?

The reasons behind this disparity need to be studied further. There may be some relationship between MCAS scores and teacher effectiveness that principals are aware of. For example, perhaps principals observe high MCAS scores from the students of teachers they consider to be effective and low MCAS scores from students of teachers they consider ineffective. If this is the case, perhaps MCAS scores are valid indicators of teacher effectiveness. This surprising finding merits further attention because principals possess valuable insight into the relationship between individual teachers and the scores of their students. Principals can and should be used as a resource to assist in decision making.

The question at the heart of this study, are MCAS scores a valid indicator of teacher effectiveness, must be addressed. Although certainly not conclusive, the results of this research indicate extreme caution should be exercised before concluding they are. As discussed in Chapter 2, The Standards (1999) list five sources of validity evidence: evidence based on test content; evidence based on response processes; evidence based on internal structure; evidence based on relations to other variables; and evidence based on consequences of testing. This study is, as you will recall from Chapter 2, a preliminary exploration of the last category, consequences of testing.

The Standards (1999) made the distinction that consequences refer specifically to evaluating whether or not the test scores are actually providing the information the proposed interpretation purports to provide. In this situation, MCAS scores are supposed to tell us if a teacher is teaching the Frameworks. The results of the analyses for research question 2 provide grounds for concern in that area. The vast majority of statistical tests do not suggest a significant relationship between the teachers' Use Scale scores and the

MCAS scores of their students. Those few tests that do have a significant relationship fail to model an adequate explanation in the variability of the MCAS scores. These preliminary finding are important in light of the advice given in the Standards that says "evidence about consequences may be directly relevant to validity when it can be traced to sources of invalidity such as construct underrepresentation or construct-irrelevant components " (p. 16). The failure of the statistical analyses in this study to link the Use Scale to the variability in test scores suggests there is another factor, or other factors, influencing test scores. This has been supported by the literature. In this particular situation, where the construct of interest is determining if a teacher is effective and defining effectiveness as teaching the Frameworks, all of the factors known to influence student test scores become construct-irrelevant variance and raise questions about the validity of using MCAS scores as indicators of teacher effectiveness.

This study has been severely limited by the inability to link individual student MCAS scores directly to each teacher. The analyses relied on self-report data, which is notoriously unreliable, and state reported school level data aggregated across grades regardless of whether one or all of the teachers in the school participated in this survey. Furthermore, the use of a survey measure and characteristics unique to this sample of respondents may have influenced the results. Future research must study the direct link between individual teacher practice related to the use of the Frameworks to plan instruction and the actual MCAS scores of the students, rather than the percent of students scoring in performance categories. Given concerns about confidentiality, this is an undertaking that the State must at the very least support, if not initiate.

The remaining four categories of validity evidence addressed in The Standards deserve special consideration here. It was not the goal of this study to isolate and explore only one category of validity evidence. It was, quite simply, not possible to explore the others. The reason for that is this: nowhere in The Standards are guidelines provided on how to use test scores from one person to make inferences about another person. All of the guidelines pertain to ensuring the development of a high quality instrument that provides accurate information about the *test taker*. The State of Massachusetts, and many others, is entering uncharted waters by using student scores as indicators of teacher effectiveness. There are not even professional guidelines available to guide the way. Yet the State has done nothing to gauge, let alone prove, that this course of action is appropriate. Given the potential this policy has to impact teacher pay and job eligibility, it is incumbent upon the State to provide definitive proof that this policy is sound and provides the information it is intended on teacher effectiveness.

Attempting to establish validity evidence based on test content is muddied here because teachers are not the test takers. We know MCAS is aligned with the Frameworks, so from a student perspective that provides validity evidence; however, looking for information about teacher effectiveness based on MCAS inserts a very complicated link in the chain: the students. Again, this is validity one step removed.

Attempting to gather validity evidence based on response processes is not a reasonable approach to this situation because there is only evidence about *student* thought processes. There is no scenario I am aware of or can think of whereby the thought processes of the students taking the test can be extended to the teachers planning instruction. It is entirely plausible to think a student can answer questions on a topic correctly based on knowledge that did not come from the classroom or that a student can completely forget a topic that was thoroughly taught. How can these threads be untangled based only on test scores?

Validity evidence based on internal structure suffers from the same limitations as the previous category of validity evidence. Methods used to gather this information rely on student responses to the test. In order for this type of evidence to be gathered, I can only think that teachers themselves would need to take a test specifically about the content of the Frameworks; it cannot be collected inferentially.

Finally, there is the category of evidence based on relations to other variables. I believe this category of validity evidence can and must be explored. Convergent validity evidence can be collected by comparing student MCAS results of a teacher's students to other measures of teacher evaluation. If it is found that the level of teacher effectiveness determined by student MCAS scores is in line with the outcomes of other methods of teacher evaluation, perhaps a relationship between teacher effectiveness and student test scores can be established. Along the same lines, predictive evidence of validity for the criterion (teacher effectiveness) could be examined in a predictive study of the relationship between MCAS scores and outcomes of other methods of teacher evaluation, or vice versa. It should be noted, however, that if the State defines teacher effectiveness as teaching the Frameworks, the other methods of teacher evaluation must focus specifically on how well or poorly the teacher performs that task. The other measures used to evaluate teachers cannot include, as they currently do, things like communication with students, leadership, or teaching methods. The comparison must be apples to apples, not apples to oranges. Otherwise, the State must acknowledge that effective teaching involves many factors beyond simply teaching the content of the Frameworks, which would raise questions about using MCAS scores as indicators of teacher effectiveness at all.

One drawback to the type of study described above is the ambiguous ethical dilemma involved in placing a teacher with poor effectiveness ratings or poor evaluations in charge of a classroom. Regardless of the design, future research into this type of validity evidence should be undertaken as it can directly compare student MCAS scores to external evaluations of teacher effectiveness.

Gathering validity evidence for this proposed new use of MCAS scores, which has not been done, will prove challenging. Traditionally, validity evidence supports test score interpretations about the test taker. Supporting test score interpretations about a third party, even an involved third party, will require innovative methods that are unknown at this point. The preliminary steps taken in this study to link the degree of use of the Frameworks to plan instruction to MCAS scores, thus hinting that MCAS scores may be a valid indicator of teacher effectiveness, were unsuccessful. This may be an artifact of data used and different results may be observed if student MCAS scores were linked directly to teachers; however, this information is controlled by the State and not easily accessible to researchers. The State is also in possession and control of teacher evaluation data that would allow some inroads to be made into exploring validity based on external factors. The responsibility for future research, which is desperately needed considering the high stakes for teachers, rests squarely on the State of Massachusetts.

There is another topic that bears consideration. The discussion to this point has focused on the complexities of establishing a (valid) relationship between student scores and teacher practice. Realistically, within this framework, it is not just the students' performance that dictates how a teacher will judged to be effective because parental factors are known to have a huge impact on student test scores. Teachers are therefore, by extension, being judged on parent performance, too. The following concerns about factors that impact student test scores are taken directly from the comment section of the survey: cognitive and social delays, behavioral issues, English as a second language, students who come to a teacher below grade level, home life, the best teachers get the most challenging students, student absences, students not getting enough sleep, students with IEPs, test anxiety, classroom temperature, students not taking medication on time, violent homes, drug-addicted parents. These are all factors that are known to influence student test scores or could conceivably influence student test scores. So, now teachers are being judged on student performance, parent performance, and community performance. This is a very slippery slope.

There were also two surprising comments that were mentioned several times in the survey comments: students do not care about MCAS until tenth grade when it counts and groups of students conspire to do poorly in order to get their teacher fired. If these anecdotal scenarios are true, this will also affect test scores. It is not reasonable to judge teacher effectiveness using a measure that is influenced by factors that are so far beyond the control of the teacher.

Unfortunately, the race for Race to the Top funds seems to have blinded certain parties. Bad policy that adversely impacts our teachers ultimately impacts the quality of education received by our students. There is as much danger in keeping poor teachers as there is in losing good teachers. And since we do not know exactly how using student

MCAS scores to judge teacher effectiveness will ultimately work out, either or both of these scenarios is possible. And, according to numerous comments from respondents, losing good teachers in the most important subjects is exactly what is going to happen. Respondents repeatedly wondered why they should continue teaching a subject where their pay and employment eligibility is tied to work other than their own. Furthermore, why would newly licensed teachers knowingly choose to go into a field where their effectiveness is judged on the performance of other people?

A positive theme in the survey comments, of both teachers and principals, was in adoption of the new Common Core State Standards (CCSS). The vast majority of the comments supported the new CCSS. Respondents seemed to think they were more manageable, included more depth of content, and allowed for deeper exploration of topic than the Math Frameworks. Future research will need to be conducted on teacher practice related to the new CCSS once they are fully adopted and implemented in the State.

This study is a very small first step in evaluating the validity of using MCAS scores as indicators of teacher effectiveness. It has shown that there is reason to believe that variations, perhaps large variations, in teacher practice exist among Massachusetts Math teachers in grades 3 -8. It has not been able to conclude that these variations are reflected in MCAS scores. The relationship between MCAS scores and the use of the Frameworks to plan instruction is tenuous. This result may come from limitations in the study or it may stem from an actual lack of a relationship between the two. Either way, because of the lack of understanding of the relationship and the high stakes decisions being made based on the presumed relationship, the State should not use MCAS scores as indicators of teacher effectiveness until the relationship has been proven. Therefore, if the results of this study are viewed as preliminary evidence to be used in building a validity argument, the results do not support the use of MCAS scores as valid indicators of teacher effectiveness.

APPENDIX A

PILOT STUDY RESPONSES

Question	Response Options				
How familiar are you with the MA Curriculum Frameworks?	Very Familiar 9	Familiar 13	Somewhat Familiar 1	Slightly Familiar 1	Not At All 0
*Do you think the Frameworks are a good representation of the material that should be taught? In other words, do you like the Frameworks?	Yes 19	No 3			
How often do you use the Frameworks to plan classroom instruction?	Very Often 5	Often 5	Sometimes 8	Rarely 6	Never
How do you use the Frameworks to plan instruction?	Follow Very Closely 6	Follow Somewhat 7	Loose Guideline 8	To Get Ideas 2	Do Not Use 1
*How often do you use the examples/practice exercises included in the Frameworks in your classroom instruction?	Very Often 1	Often 3	Sometimes 3	Rarely 5	Never 11
*How often do you include topics/materials in instruction that are not included in the Frameworks?	Very Often 8	Often 5	Sometimes 7	Rarely 2	Never 1
*How often do you review the released MCAS items posted by MA DESE?	Very Often 2	Often 5	Sometimes 9	Rarely 2	Never 5
*How often do you use the released MCAS items to plan instruction?	Very Often 2	Often 6	Sometimes 7	Rarely 3	Never 5
*How often do you use the released MCAS items in the classroom?	Very Often 2	Often 5	Sometimes 7	Rarely 3	Never 5
How do you use the released MCAS items? (check all that apply) To plan instruction As examples during instruction As a pretest	 12 12 6	Responses provided when *Other* selected: Test Prep Training To assist students who need to retake it			

As homework As a test To see what my students know Other (please describe)	4 5 9 6	Review			
*How often do you use the MCAS score reports provided by the MA DESE?	Very Often 1	Often 8	Sometimes 5	Rarely 8	Never 1
*To what extent do you understand the MCAS score reports?	Very well 4	Well 9	Somewhat 7	Not Well 1	Not At All 2
*How often do you find the MCAS score reports helpful?	Very Often 1	Often 2	Sometimes 12	Rarely 3	Never 5
*How often do the MCAS score reports inform your instruction?	Very Often 1	Often 3	Sometimes 10	Rarely 2	Never 6
*How often do the MCAS results agree with your personal knowledge of your students?	Very Often 3	Often 8	Sometimes 5	Rarely 2	Never 1
*Do you ever feel the Frameworks are overwhelming?	Very Often 1	Often 2	Sometimes 11	Rarely 4	Never 3
*Would you like help in using the Frameworks to guide your instruction?	Yes 6		No 17		
What do you think would be helpful in assisting teachers to use the Frameworks to guide instruction? (please check all that apply) A college course (before you begin teaching) Professional development An alignment tool An online tutorial from the MA DESE Other (please describe)	 4 13 11 6 4	Responses provided when *Other* selected: Mentoring from veteran teacher A usable website Mentors			
Open-Ended Responses					
Do you think the Frameworks are a good representation of the material that should be taught? In other words, do you like the Frameworks? If you selected "no" above, please briefly explain your answer. forcing students to study materials in Soc-Studies that are not as relevant or important to them.					

some things emphasized too much, others not enough

Frameworks do not cover upper level physics, so N/A

No framework exists for my class "Physical Science". We have looked at combination of Chemistry + Physics frameworks to develop the curriculum. We are now also committing to preparing our students for the Physics MCAS (w/o dropping the chemistry portion of the course)

More freedom would be better

If you use topics/materials in instruction that are not included in the Frameworks, please tell me why.

AB Calculus-"outside frameworks"

Because they are lovely

cross-curriculum/inter-disciplinary goals

fun + informative

high school Frameworks for physics go to grade 10- I teach 11+12, Honors/AP

I like to include current events and practical "real life" examples in my class instruction

I still think they are important-high interest keeps kids motivated

I teach an AP course

I teach social skills on a daily basis (anger/stress management, coping skills, etc)

I think it is material worth covering

I work with students w/ spec needs who have completed their requirements to receive a certificate of attendance (18-22 y.o.)

if students have taken MCAS/all frameworks addressed, delve deeper or something related to a topic of interest w/in frameworks

Supplemental materials

often times they are more creative and fit my needs better

Psychology is not part of the frameworks

support other ideas included

The Frameworks do not go far enough

To put science in a meaningful, historical + philosophical context. To add interest excitement and meaning to the curriculum. To broaden the mental horizons of students.

Who do you believe developed the Frameworks?

A state-developed committee

Academics & Educators

Educators & politicians

experts/educators/in my field

I'd like to think educators but believe it was admin level folks
i'm not sure
ivory tower non-teachers
MA state teachers
no idea
other teachers
the state
teachers
Teachers
teachers + admin
Teachers and administrators as well as government officials
teachers from around the state?
the state-dept of education
unsure
various educators, administrators, pta?
you did
Why do you believe the Frameworks were developed?
accountability
accountability, unification
all students have equal/comparable knowledge base; to prevent teaching of outlandish topics
America was "lagging behind" other countries, mirrors Japan
b/c of the ten scariest words in the English language:"Hello, I'm from the gov't, and I'm here to help"
continuity, accountability
for consistency across the state
in order to ensure students are learning approved curriculum + teachers have a map from which to develop their lessons
Money
no idea
people needed a frame
standardize curriculum
to assist with curricular development
to align Massachusetts math curriculum across the state
to create consistency throughout classrooms
to develop consistent rich content and skills
to ensure a level of proficiency is met by students

to establish a uniform set of info to be taught to students-they should all be exposed to similar info @ basically same tx
to give teachers a guideline
to keep consistency across all departments
To provide guidance to school districts on the minimum standards to adhere to.
to provide teachers guidance in providing instruction for students in areas of importance+ to ensure that all students were taught the same core info
To structuralize assessment
How often do the MCAS results agree with your personal knowledge of your students? Please explain the answer you gave to the question above:
I don't have knowledge of how individuals scored just in aggregate
I often can predict which questions students will have trouble with
I teach foreign language
I work with students w/ spec needs w/ intellectual challenges
lets me know areas they skipped/had difficulty with
my students have already taken MCAS
my students tend to score lower than their average daily performance
no MCAS music
scores match often with abilities demonstrated by students
strong students in math class (good grades) generally score well
Students taking the physics MCAS have not had a full year of physics instruction, certain topics are not covered or are covered superficially
usually meet expectations of performance
vague, they are just practice
Wording of questions sometimes affect student understanding of what is being asked
As a teacher, do you feel that you receive adequate support from administration? Please explain.
a lot of support but little time or money
building-yes, system-no
Generally, yes
No
sometimes- often "what you say"+"what you do" can turn into doing nothing.
sometimes, although when new technology was introduced we were not trained
Sometimes. Usually the limitations seem to come from on high, so it's hard to say.
usually
yes
Yes
Yes
yes fortunate to have a very supportive principal who build team spirit

yes yes yes
yes-but they are often too focused on MCAS
yes, I have been given a great deal of latitude in developing my program
yes, PDPs opportunity in house
yes, sent me to this moodle workshop
yes, very available
yes, very supportive
yes, when I need help they give it.
yes!!
As a teacher, do you feel that you have access to the resources you need? Please explain.
can always use more!
no-budget
no-money is a huge issue.
no, budget limits on purchasing technology
no, never enough $, space, stuff
No. I would use a SmartBoard.
not always for my subject matter
Often, no. Technology is very limited in availability.
sometimes - depending on availability of funds & new development in technology
Sometimes. Realistic $ issues prevent from having everything I'd like, but I can do my job well
time is lacking
usually
No. We are not allowed to install software and that makes it difficult to try new stuff.
yes
yes
yes
yes
yes supportive admin
yes, resources are out there
yes, very supportive
yes!!

APPENDIX B

INTRODUCTORY EMAIL TO TEACHERS

Dear Teacher,

I am a doctoral candidate in the Psychometric Methods, Educational Statistics, and Research Methods (PMEDRM) Program in the School of Education at the University of Massachusetts Amherst. My dissertation focuses on evaluating whether MCAS scores are valid indicators of teacher effectiveness. This topic is of great and immediate concern because the Massachusetts Department of Elementary and Secondary Education recently announced that MCAS scores, when available, are to be used as a significant indicator of teacher effectiveness.

Currently, there is no research available to support the use of MCAS scores for this purpose. I believe that if Massachusetts is to use MCAS scores as an indicator of teacher effectiveness, the validity of the interpretations made about teacher effectiveness based on MCAS scores must be evaluated. To that end, I am surveying Massachusetts public school principals and math teachers in grades three through eight about their use of the Curriculum Frameworks, their opinions about the Frameworks, the performance of their students on MCAS, their experiences in teacher preparation programs and with professional development, and how they think teacher effectiveness should be evaluated.

This survey should take 10 – 15 minutes to complete. I recognize that you are very busy with the job of educating our children so I am offering an incentive to those who respond. Each person who responds to the survey, and provides an email address, will be entered into a drawing. Ten names will be randomly drawn and each winner will receive a $100.00 Amazon.com gift card. Respondents may choose to remain completely anonymous; however, you will not be eligible to be entered into the drawing because there will be no way to notify you if you win. The drawing will occur in July, so please provide an email address that you will be checking at that time. Please note that this survey will be open for two weeks. If you are interested in responding, please do so before June 22, 2012 – this will also ensure you are entered into the drawing to win a $100.00 Amazon.com gift card!

I will be happy to answer any questions you may have. Please do not hesitate to contact me at jcopella@educ.umass.edu if I can provide any further information or answer any questions. You may also contact my advisor at UMASS, Dr. Lisa Keller, at lkeller@educ.umass.edu with any questions or concerns.

Your honest and anonymous answers to these survey questions will be an invaluable resource for evaluating the appropriateness of using MCAS scores to evaluate teacher effectiveness. If you are willing to share your thoughts and experiences, please click on the link below:

http://www.surveymonkey.com/s/DSVQ52Q

Thank you for your time and consideration.

Best regards
Jenna Copella

APPENDIX C

INTRODUCTORY EMAIL TO PRINCIPALS

Dear Principal,

I am a doctoral candidate in the Psychometric Methods, Educational Statistics, and Research Methods (PMESRM) Program in the School of Education at the University of Massachusetts Amherst. My dissertation focuses on evaluating whether MCAS scores are valid indicators of teacher effectiveness. This topic is of great and immediate concern because the Massachusetts Department of Elementary and Secondary Education recently announced that MCAS scores, when available, are to be used as a significant indicator of teacher effectiveness.

Currently, there is no research available to support the use of MCAS scores for this purpose. I believe that if Massachusetts is to use MCAS scores as an indicator of teacher effectiveness, the validity of the interpretations made about teacher effectiveness based on MCAS scores must be evaluated. To that end, I am surveying Massachusetts public school principals and math teachers in grades three through eight about their use of the Curriculum Frameworks, their opinions about the Frameworks, the performance of their students on MCAS, their experiences in teacher preparation programs and with professional development, and how they think teacher effectiveness should be evaluated.

This survey should take 10 – 15 minutes to complete. I recognize that you are very busy with the job of educating our children so I am offering an incentive to those who respond. Each person who responds to the survey, and provides an email address, will be entered into a drawing. Ten names will be randomly drawn and each winner will receive a $100.00 Amazon.com gift card. Respondents may choose to remain completely anonymous; however, you will not be eligible to be entered into the drawing because there will be no way to notify you if you win. The drawing will occur in July, so please provide an email address that you will be checking at that time. Please note that this survey will be open for two weeks. If you are interested in responding, please do so before June 22, 2012 – this will also ensure you are entered into the drawing to win a $100.00 Amazon.com gift card!

I will be happy to answer any questions you may have. Please do not hesitate to contact me at jcopella@educ.umass.edu if I can provide any further information or answer any questions. You may also contact my advisor at UMASS, Dr. Lisa Keller, at lkeller@educ.umass.edu with any questions or concerns.

Your honest and anonymous answers to these survey questions will be an invaluable resource for evaluating the appropriateness of using MCAS scores to evaluate teacher effectiveness. If you are willing to share your thoughts and experiences, please click on the link below:

http://www.surveymonkey.com/s/DWM3R2Y

Thank you for your time and consideration.
Best regards
Jenna Copella

APPENDIX D

TEACHER REMINDER EMAIL

Dear Teacher,

A couple of weeks ago you received an email describing my dissertation research, which involves evaluating the validity of MCAS scores as indicators of teacher effectiveness. The email also contained a link to a survey. I am writing again to ask you to please consider participating in this research if you have not done so already. Each Teacher response provides badly needed insight into this issue and strengthens the results of this research.

The survey is intended for Massachusetts public school principals and math teachers in grades three through eight. The questions ask about their use of the Curriculum Frameworks, their opinions about the Frameworks, the performance of their students on MCAS, their experiences in teacher preparation programs and with professional development, and how they think teacher effectiveness should be evaluated. All responses will be kept strictly confidential and will not be reported in any way that would allow identification of individuals or schools. The survey should take you only 10 – 15 minutes to complete.

Each person who responds to the survey, and provides an email address, will be entered into a drawing to win one of ten $100.00 Amazon.com gift cards. The ten winners will be randomly drawn. Respondents may choose to remain completely anonymous; however, you will not be eligible to be entered into the drawing because there will be no way to notify you if you win. The drawing will occur in July, so please provide an email address that you will be checking at that time. Please note that this survey will only be open until June 22, 2012. If you are interested in responding, please do it soon – this will also ensure you are entered into the drawing to win a $100.00 Amazon.com gift card!

I will be happy to answer any questions you may have. Please do not hesitate to contact me at jcopella@educ.umass.edu if I can provide any further information or answer any questions. You may also contact my advisor at UMASS, Dr. Lisa Keller, at lkeller@educ.umass.edu with any questions or concerns.

Don't miss this chance to contribute to educational research and to be entered into a drawing to win a $100.00 Amazon.com gift card! Please follow the link below to the survey:

http://www.surveymonkey.com/s/DSVQ52Q

Best regards
Jenna Copella

APPENDIX E

PRINCIPAL REMINDER EMAIL

Dear Principal,

A couple of weeks ago you received an email describing my dissertation research, which involves evaluating the validity of MCAS scores as indicators of teacher effectiveness. The email also contained a link to a survey. I am writing again to ask you to please consider participating in this research if you have not done so already. Each Principal response provides badly needed insight into this issue and strengthens the results of this research.

The survey is intended for Massachusetts public school principals and math teachers in grades three through eight. The questions ask about their use of the Curriculum Frameworks, their opinions about the Frameworks, the performance of their students on MCAS, their experiences in teacher preparation programs and with professional development, and how they think teacher effectiveness should be evaluated. All responses will be kept strictly confidential and will not be reported in any way that would allow identification of individuals or schools. The survey should take you only 10 – 15 minutes to complete.

Each person who responds to the survey, and provides an email address, will be entered into a drawing to win one of ten $100.00 Amazon.com gift cards. The ten winners will be randomly drawn. Respondents may choose to remain completely anonymous; however, you will not be eligible to be entered into the drawing because there will be no way to notify you if you win. The drawing will occur in July, so please provide an email address that you will be checking at that time. Please note that this survey will only be open until June 22, 2012. If you are interested in responding, please do it soon – this will also ensure you are entered into the drawing to win a $100.00 Amazon.com gift card!

I will be happy to answer any questions you may have. Please do not hesitate to contact me at jcopella@educ.umass.edu if I can provide any further information or answer any questions. You may also contact my advisor at UMASS, Dr. Lisa Keller, at lkeller@educ.umass.edu with any questions or concerns.

Don’t miss this chance to contribute to educational research and to be entered into a drawing to win a $100.00 Amazon.com gift card! Please follow the link below to the survey:

http://www.surveymonkey.com/s/DWM3R2Y

Best regards
Jenna Copella

APPENDIX F

TEACHER SURVEY

1. Is your gender:			
Total	Female	Male	Omitted
745	84%(624)	15%(112)	1%(9)

2. What is the name of the school you currently teach in?	
Responses	Omitted
745	0

3. How long have you been teaching?	
Mean (years)	Standard Deviation (years)
14.1	9.1

4. How long have you been teaching at this school?	
Mean (years)	Standard Deviation (years)
8.8	7.2

5. What grade level(s) do you teach (please check all that apply)?					
3	**4**	**5**	**6**	**7**	**8**
24%(178)	27%(202)	27%(199)	19%(141)	15%(111)	15%(111)

Listed below are statements related to the Massachusetts Curriculum Frameworks in Math (the Math Frameworks). Please read each statement and indicate the degree to which you agree with the statement on the rating scale provided.						
	Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree	Response Count
6. I am familiar with the Math Frameworks.	1%(8)	0%(3)	3%(18)	34%(236)	61%(420)	92%(685)
7. The Math Frameworks are helpful for instructional planning.	1%(6)	3%(18)	8%(55)	51%(350)	37%(257)	92%(686)
8. The Math Frameworks are a good representation of the material that should be taught.	1%(4)	8%(53)	16%(110)	55%(373)	21%(143)	92%(683)
9. All of the content included in the Math Frameworks is equally important.	4%(27)	40%(273)	20%(136)	28%(191)	8%(58)	92%(685)
10. I always use the Math Frameworks when developing a lesson plan.	3%(23)	17%(115)	24%(161)	36%(247)	20%(135)	91%(681)
11. I use the Math Frameworks mostly as a loose guideline when planning instruction.	9%(60)	28%(188)	23%(154)	37%(249)	5%(31)	92%(682)
12. The content of the Math Frameworks appropriately represents what students need to learn.	2%(12)	17%(118)	23%(159)	50%(338)	8%(55)	92%(682)
13. The Math Frameworks try to cover too much material.	2%(14)	16%(110)	22%(149)	34%(234)	25%(172)	91%(679)
14. There is not enough instructional time to cover the content of the Math Frameworks.	2%(14)	13%(86)	16%(107)	37%(248)	33%(222)	91%(677)
15. The Math Frameworks allow for in-depth coverage of content.	13%(92)	42%(287)	24%(162)	17%(115)	4%(24)	91%(680)
16. The Math Frameworks are missing important	4%(25)	32%(215)	43%(288)	18%(124)	4%(24)	91%(676)

content.						
17. I follow the Math Frameworks closely when I plan instruction.	4%(25)	14%(97)	24%(161)	42%(284)	16%(110)	91%(677)
18. The Math Frameworks are overwhelming.	5%(33)	27%(179)	33%(222)	27%(178)	9%(59)	90%(671)
19. I would like help with using the Math Frameworks to plan instruction.	14%(92)	38%(258)	26%(176)	18%(125)	4%(25)	91%(676)
20. I have no interest in using the Math Frameworks for instructional planning.	47%(316)	37%(250)	13%(90)	2%(15)	1%(8)	91%(679)
21. Teachers should have complete autonomy when deciding what to teach.	42%(286)	41%(279)	12%(81)	4%(25)	2%(12)	92%(683)

22. My teacher preparation program adequately prepared me to plan instruction using the Math Frameworks.		
	Response Percent	Response Count
Strongly Disagree	10.7%	73
Disagree	23.8%	163
Neither Agree or Disagree	22.0%	150
Agree	26.8%	182
Strongly Agree	12.0%	83
I have not attended a teacher prep program	4.6%	29
	Responses	680
	Omitted	66

23. Have you ever attended a Professional Development designed to help teachers use the Math Frameworks to plan instruction? This question is concerned with any Professional Development that attempted to help teachers directly link the content of their lessons to the content of the Math Frameworks. The Professional Development may have been focused on a specific program (such as Backward Design or another curriculum mapping strategy) or it may have been about a less formal method.		
	Response Percent	Response Count
Yes	68.4%	463
No	31.6%	214
	Responses	677
	Omitted	69

24. Approximately how many Professional Development(s) to help teachers use the Math Frameworks to plan instruction have you attended?	
0	2
0.5	1
1	53
2	85
3	68
4	51
5	61
6	20
7	2
8	9
9	1
10	30
12	3
15	5
18	1
20	9
25	2
30	2
50	1
60	1
Non-numeric responses indicating attendance	54
TOTAL	62%(461)
Omitted	38%(285)

25. Were the Professional Development(s) you attended helpful in assisting you to use the Math Frameworks to plan instruction?			
Not at all Helpful	**Somewhat Helpful**	**Helpful**	**Very helpful**
5%(24)	54%(252)	31%(144)	9%(43)
		Total Responses	62%(463)
		Omitted	38%(283)

Listed below are statements related to ways teachers may use the Math Frameworks. Please read each statement and, using the rating scale provided, indicate the degree to which the statement reflects your use of the Math Frameworks.						
	Never	Rarely	Sometimes	Often	Always	Response Count
26. I use the Math Frameworks to plan instruction.	4% (28)	6% (37)	24% (158)	37% (241)	29% (188)	88% (652)
27. I refer back to the Math Frameworks for guidance during the course of planning instruction.	5% (33)	8% (55)	28% (179)	37% (241)	22% (142)	87% (650)
28. I include topics in instruction that are NOT included in the Math Frameworks.	6% (37)	23% (150)	50% (329)	17% (110)	4% (27)	88% (653)
29. I use a program that facilitates the use of the Math Frameworks to plan instruction (e.g., curriculum mapping or	14% (88)	12% (81)	19% (121)	35% (229)	20% (132)	87% (651)

some other program).						
30. When planning instruction I create entire lessons with content that does NOT appear in the Math Frameworks.	36% (236)	44% (289)	15% (99)	3% (17)	2% (10)	87% (651)
31. I provide my students with more information (theories, formulas, steps to follow, “tricks”, etc.) than is listed in the Math Frameworks.	4% (27)	7% (43)	30% (193)	44% (287)	16% (101)	87% (651)
32. I use information received from professional development to plan instruction using the Math Frameworks.	11% (73)	14% (93)	34% (224)	33% (216)	7% (45)	87% (651)
33. My colleagues and I talk about ways to use the Math Frameworks in instructional planning.	9% (59)	18% (117)	28% (181)	32% (210)	13% (85)	88% (652)

Listed below are statements related to ways teachers may use MCAS Math results. Please read each statement and, using the rating scale provided, indicate the degree to which the statement reflects your use of the MCAS Math results.						
	Never	Rarely	Sometimes	Often	Always	Response Count
34. I am eager to get my students’ MCAS Math results.	4% (26)	6% (42)	14% (93)	27% (175)	48% (315)	87% (651)
35. I pretty much ignore my students’ MCAS Math results.	66% (433)	21% (136)	9% (59)	3% (19)	1% (5)	88% (652)
36. I spend a lot of time looking at MCAS results soon after I receive them.	2% (16)	9% (60)	20% (127)	31% (202)	38% (246)	87% (651)
37. After I get my students’ MCAS results, I evaluate student performance with respect to the content strands in the Math Frameworks.	6% (36)	10% (68)	22% (142)	29% (190)	33% (214)	87% (650)

Listed below are activities that may be helpful in assisting teachers to plan instruction based on the Math Frameworks. Please use the scale provided to indicate the degree to which you believe each activity would be helpful in assisting teachers to plan instruction based on the Math Frameworks.					
	Not helpful at all	**Minimally helpful**	**Helpful**	**Very helpful**	**Response Count**
38. A college course (before you begin teaching)	6% (36)	22% (144)	42% (273)	30% (192)	87% (645)
39. Professional development	1% (7)	8% (53)	49% (314)	42% (273)	87% (647)
40. An alignment tool	1% (7)	11% (73)	45% (291)	42% (272)	86% (643)
41. An online tutorial	13% (82)	36% (233)	36% (232)	15% (98)	87% (645)

from the MA DESE					
42. Collaboration with other teachers	0% (3)	2% (15)	25% (159)	73% (469)	87% (646)

	Mean (%)	Standard Deviation (%)	Response Count	Omit
43. What percent of your students do you estimate scores in the Needs Improvement range on MCAS Math last year?	24.4	19.2	620	126
44. What percent of your students do you estimate scored in the Proficient range on MCAS Math last year?	51.6	21	619	127
45. What percent of your students do you estimate scored in the Advanced range on MCAS Math last year?	18	14.6	613	133

Listed below are activities that have been suggested as possible ways to evaluate teacher effectiveness. Please read each activity and, using the rating scale provided, indicate the degree to which you believe the activity is a useful way to evaluate teacher effectiveness.					
	Not at all useful	Minimally useful	Useful	Very useful	Response Count
46. Peer observation	6% (35)	17% (108)	51% (321)	26% (167)	85% (631)
47. Principal observation	4% (25)	24% (151)	57% (361)	15% (97)	85% (634)
48. Outside evaluator	13% (81)	38% (240)	41% (259)	8% (52)	85% (632)
49. Test scores	16% (104)	47% (300)	32% (205)	4% (23)	85% (632)
50. Parent evaluations	31% (195)	50% (314)	19% (120)	1% (5)	85% (634)
51. Student evaluations	17% (108)	41% (256)	35% (223)	7% (45)	85% (632)
52. Videotaping the teacher during instruction	13% (80)	29% (182)	43% (273)	15% (94)	84% (629)

53. Is there a person in your school (not necessarily a curriculum director) to help teachers with using the Frameworks to plan instruction?		
	Percent	Response Count
Yes	51	323
No	49	312
	Total Responses	635
	Omitted	111

54. Do you teach English Language Arts as well as Math?		
	Percent	Response Count
Yes	63	401
No	37	232
	Total Responses	633
	Omitted	113

55. How do you use the ELA Frameworks to plan instruction?		
	Response Percent	Response Count
About the same amount as the Math Framework	68.8%	275
More than I use the Math Frameworks	17.5%	70
Less than I use the Math Frameworks	12.5%	50
I don't use the ELA Frameworks to plan instruction	1.3%	5
	Total	400

APPENDIX G

PRINCIPAL SURVEY

1. What is the name of the school you currently teach in?	
Responses	Omitted
147	9

2. Is your gender:			
Female	Male	Responses	Omitted
65% (102)	33% (51)	98% (153)	2% (3)

3. How long have you been a principal at this school?			
Mean (yrs)	Standard Dev. (yrs)	Responses	Omitted
5.4	4.2	153	3

4. Were you ever a classroom teacher? (routing rule question)			
Yes	No	Responses	Omitted
92% (144)	6% (10)	99% (154)	1% (2)

5. How long were you a classroom teacher?			
Mean (yrs)	Standard Dev. (yrs)	Responses	Omitted
12.4	6.8	145	11

6. What percent of the students in your school qualify for free or reduced lunch?			
Mean (%)	Standard Dev. (%)	Responses	Omitted
35	29	144	12

Listed below are statements related to the Massachusetts Curriculum Frameworks in Math (the Math Frameworks). Please read each statement and indicate the degree to which you agree with the statement on the rating scale provided.						
	Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree	Response Count
8. I am familiar with the Math Frameworks.	0% (0)	0% (0)	2% (3)	61% (83)	37% (51)	88% (137)
9. The Math Frameworks are a good representation of the material that should be taught.	0% (0)	2% (3)	6% (8)	75% (103)	17% (23)	88% (137)
10. All of the content in the Math Frameworks is equally important.	1% (1)	42% (57)	18% (24)	36% (49)	4% (6)	88% (137)

11. The content of the Math Frameworks appropriately represents what students need to learn.	0% (0)	9% (12)	20% (27)	64% (87)	8% (11)	88% (137)
12. The Math Frameworks try to cover too much material.	1% (1)	25% (34)	27% (37)	31% (42)	16% (22)	87% (136)
13. There is not enough instructional time to cover the content of the Math Frameworks.	2% (3)	19% (26)	21% (29)	35% (48)	22% (30)	87% (136)
14. The Math Frameworks allow for in-depth coverage of content.	6% (8)	35% (47)	23% (31)	34% (46)	3% (4)	87% (136)
15. The Math Frameworks are missing important content.	3% (4)	37% (49)	47% (63)	11% (15)	2% (2)	85% (133)
16. The Frameworks are overwhelming.	5% (7)	34% (46)	30% (40)	25% (34)	5% (7)	86% (134)
17. Teachers should have complete autonomy when deciding what to teach.	50% (67)	45% (60)	1% (2)	3% (4)	1% (1)	86% (134)

Listed below are statements related to ways that you and the teachers in your school may use the Math Frameworks. Please read each statement and, using the rating scale provided, indicate the degree to which the statement reflects the use of the Math Frameworks in your school.

	Never	Rarely	Sometimes	Often	Always	Response Count
18. In my school, we have a program (Backward design, mapping, etc.) that helps the teachers use the Math Frameworks to plan instruction.	5%(6)	12%(15)	29%(37)	35%(44)	20%(25)	81%(127)
19. The teachers in my school use the approved program plan instruction using the Math Frameworks.	5% (6)	4% (5)	16% (20)	55% (70)	20% (26)	81% (127)
20. In my school, teachers include topics in instruction that are NOT included in the Math Frameworks.	6% (7)	36% (46)	43% (55)	13% (16)	2% (3)	81% (127)
21. In my school, teachers create entire lessons with content that does NOT appear in the Math Frameworks.	26% (33)	52% (66)	17% (22)	3% (4)	2% (2)	81% (127)
22. In my school, teachers provide students with more information (theories, formulas, steps to follow,	2% (3)	14% (18)	43% (55)	35% (45)	5% (6)	81% (127)

"tricks", etc.) than is listed in the Math Frameworks.						
23. The teachers in my school and I talk about ways to use the Math Frameworks in instructional planning.	2% (2)	6% (7)	31% (39)	50% (64)	12% (15)	81% (127)

Listed below are statements related to ways you may use MCAS Math results. Please read each statement and, using the rating scale provided, indicate the degree to which the statement reflects your use of the MCAS Math results.						
	Never	Rarely	Sometimes	Often	Always	Response Count
24. I spend a lot of time looking at MCAS Math results soon after I receive them.	0% (0)	2% (2)	9% (11)	38% (49)	52% (66)	82% (128)
25. After I get MCAS Math results for the school, I evaluate student performance with respect to the content strands in the Math Frameworks.	2% (2)	2% (2)	8% (10)	40% (51)	49% (62)	81% (127)

Listed below are activities that may be helpful in assisting teachers to plan instruction based on the Math Frameworks. Please use the scale provided to indicate the degree to which you believe each activity would be helpful in assisting teachers to plan instruction based on the Math Frameworks.					
	Not helpful at all	Minimally helpful	Helpful	Very helpful	Response Count
26. A college course (before you begin teaching)	4%(5)	21%(27)	47%(60)	28%(35)	81%(127)
27. Professional development	0%(0)	2%(2)	41%(52)	57%(73)	81%(127)
28. An alignment tool	1%(1)	5%(6)	55%(70)	39%(50)	81%(127)
29. An online tutorial from the MA DESE	10%(12)	37%(46)	43%(54)	11%(14)	81%(126)
30. Collaboration with other teachers	0%(0)	0%(0)	19%(24)	81%(103)	81%(127)

Listed below are activities that have been suggested as possible ways to evaluate teacher effectiveness. Please read each activity and, using the rating scale provided, indicate the degree to which you believe the activity is a useful way to evaluate teacher effectiveness.					
	Not at all useful	Minimally useful	Useful	Very useful	Response Count
31. Peer observation	0%(0)	13%(17)	55%(69)	32%(40)	81%(126)
32. Principal observation	0%(0)	4%(5)	53%(67)	43%(54)	81%(126)
33. Outside evaluator	4%(5)	33%(41)	54%(67)	10%(12)	80%(125)
34. Test scores	2%(3)	27%(34)	61%(76)	10%(12)	80%(125)
35. Parent evaluations	13%(16)	62%(78)	24%(30)	2%(2)	81%(126)
36. Student evaluations	4%(5)	38%(48)	48%(60)	10%(12)	80%(125)
37. Videotaping the teacher during instruction	1%(1)	13%(16)	46%(57)	41%(51)	80%(125)

38. Is there a person in your school (not necessarily a curriculum director) to help teachers with using the Frameworks to plan instruction?		
	Response Percent	Response Count
Yes	67.5%	85
No	32.5%	41

APPENDIX H

FACTOR LOADINGS

Survey questions	Factors	
	1	2
6. I am familiar with the Math Frameworks.	**.797**	.347
7. The Math Frameworks are helpful for instructional planning.	**.800**	.374
8. The Math Frameworks are a good representation of the material that should be taught.	**.788**	.382
9. All of the content included in Math Frameworks is equally important.	**.780**	.397
10. I always use the Math Frameworks when developing a lesson plan.	**.733**	.424
11. I use the Math Frameworks mostly as a loose guideline when planning instruction.	**.760**	.320
12. The content of the Math Frameworks appropriately represents what students need to learn.	**.797**	.361
13. The Math Frameworks try to cover too much material.	**.785**	.269
14. There is not enough instructional time to cover the content of the Math Frameworks.	**.780**	.252
15. The Math Frameworks allow for in-depth coverage of content.	**.753**	.417
16. The Math Frameworks are missing important content.	**.746**	.413
17. I follow the Math Frameworks closely when I plan instruction.	**.725**	.462
18. The Math Frameworks are overwhelming.	**.780**	.265
19. I would like help with using the Math Frameworks to plan instruction.	**.772**	.369
20. I have no interest in using the Math Frameworks for instructional planning.	**.790**	.383
21. Teachers should have complete autonomy when deciding what to teach.	**.816**	.391
26. I use the Math Frameworks to plan instruction.	.364	**.865**
27. I refer back to the Math Frameworks for guidance during the course of planning instruction.	.358	**.860**
28. I include topics in instruction that are NOT included in the Math Frameworks.	.397	**.829**
29. I use a program that facilitates the use of the Math Frameworks to plan instruction (e.g., curriculum mapping or some other program).	.337	**.814**
30. When planning instruction I create entire lessons with content that does NOT appear in the Math Frameworks.	.416	**.798**
31. I provide my students with more information (theories, formulas, steps to follow, "tricks", etc.) than is listed in the Math Frameworks.	.376	**.819**
32. I use information received from professional development to plan instruction using the Math Frameworks.	.368	**.852**
33. My colleagues and I talk about ways to use the Math Frameworks in instructional planning.	.360	**.858**

BIBLIOGRAPHY

Aguirre-Munoz, Z. & Boscardin, C.K. (2008). Opportunity to learn and English learner achievement: Is increased content exposure beneficial? *Journal of Latinos and Education, 7*(3), 186-205.

Airasian, P.W. & Miranda, H. (2002). The role of assessment in the revised taxonomy. *Theory Into Practice, 41*(4), 249-254.

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

American Federation of Teachers (2009). The instructional demands of standards-based reform (Item No. 39-0186). Washington, DC.

Ananda, S. (2003). Achieving alignment. *Leadership, 33*(1), 18-22.

Anderson, L.W. (2002). Curricular alignment: A re-examination. *Theory into Practice, 41*(4), 255-260.

Betebenner, D.W. (2009). Growth, standards, and accountability. Colorado Department of Education Technical Report.

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

Bober, M.J., Sullivan, H.J., Lowther, D.L., & Harrison, P. (1998). Instructional practices of teachers enrolled in educational technology and general education programs. *Educational Technology Research & Development, 46*(3), 81-97.

Boscardin, C.K., Aguirre-Munoz, Z., Stoker, G., Kim, J., Kim, M., & Lee, J. (2005). Relationship between opportunity to learn and student performance on English and Algebra assessments. *Educational Assessment, 10*(4), 307-332.

Braun, H. (2004). *Value-added modeling: What does due diligence require?* Educational TestingService. Retrieved on July 1, 2011 from http://www.cgp.upenn.edu/pdf/Braun%20-%20VA%20Modeling%20What%20Does%20Due%20Diligence%20Req.pdf

Chester, M. (2010). Letter to educators explaining the growth model. Retrieved on December 27, 2011 from http://www.doe.mass.edu/mcas/growth/0910letter.html .

Childre, A., Sands, J.R., & Pope, S.T. (2009). Backward design: Targeting depth of understanding for all learners. *Teaching Exceptional Children*, *41*(5), 6-14.

Cooley, W.W. & Leinhardt, G. (1978). *The instructional dimensions study: The search for effective classroom processes.* Pittsburgh, PA: Pittsburgh University Learning Research and Development Center.

Cohen, S.A. (1987). Instructional alignment: Searching for the magic bullet. *Educational Researcher*, *16*, 16-20.

Elia, J.I. (1994). An alignment/transfer experiment with low socioeconomic level students. Teacher Education Quarterly, 21(3), 113-124.

Gravetter, F.J. & Wallnau, L.B. (2008). Essentials of statistics for the behavioral sciences. 6th Edition. Belmont, CA: Thomson Wadsworth.

Hamilton, L.S., Stecher, B.M., & Yuan, K. (2009). Standards-Based reform in the United States: History, research, and future directions. Washington, DC: Center on Education Policy.

Hamm, J.D. (1994). Drawing a new picture: Describing the work of exemplary curriculum directors. Paper presented at the annual meeting of the Washington State Association of Supervisors and Curriculum Development, Spokane, WA, February 3-5, 1994.

Herman, J.L. & Webb, N.M. (2007). Alignment methodologies. *Applied Measurement in Education*, *20*(1), 1-5.

Hinchey, P. H. (2010). *Getting Teacher Assessment Right: What Policymakers Can Learn from Research*. Boulder, CO: National Education Policy Center. Retrieved November 27, 2011 from http://nepc.colorado.edu/publication/getting-teacher-assessment-right.

Jacobs, B. (2000). Implementing Standards: The California Mathematics Textbook Debacle. *The Phi Delta Kappan, 83*(3), 264-272.

James, A.R., Griffin, L.L., & Dodds, P. (2008). The relationship between instructional alignment and the ecology of physical education. *Journal of Teaching in Physical Education*, *27*, 308-326.

Kane, M. T. (1990). An argument-based approach to validation. ACT Research Report Series 90-13. Iowa City, IA: American College Testing Program.

Kane, M. T. (2006). Validation. In R.L. Brennan (Ed.), *Educational Measurement* (4th Ed.). Westport, CT: American Council on Education/Prager.

Kupermintz, H. (2003). Teacher Effects and Teacher Effectiveness: A Validity Investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis, 25,* 287-298.

Leinhardt, G. & Seewald, A.M. (1981). What's tested, what's taught? *Journal of Educational Measurement, 18*(2), 85-96.

Levine, D.U. (1982). Successful approaches for improving academic achievement in inner city elementary schools. *The Phi Delta Kappan, 63*(8), 523-526.

Martone, A. & Sireci, S.G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79*(4), 1332-1361.

Massachusetts Department of Elementary and Secondary Education (2008). Ensuring technical quality: Policies and procedures guiding the development of the MCAS tests. Malden, MA: Massachusetts Department of Elementary and Secondary Education.

Massachusetts Department of Elementary and Secondary Education (2010). 2009 MCAS technical report. Malden, MA: Massachusetts Department of Elementary and Secondary Education.

Massachusetts Department of Elementary and Secondary Education (2011a). Building a breakthrough framework for educators in the Commonwealth. Malden, MA: Department of Elementary and Secondary Education.

Massachusetts Department of Elementary and Secondary Education (2011b). Final regulations on evaluation of educators. Retrieved on December 27, 2011 from http://www.doe.mass.edu/lawsregs/603cmr35.html?section=all .

Massachusetts Department of Elementary and Secondary Education (2011c). Retrieved on December 27, 2011 from http://profiles.doe.mass.edu/profiles/teacher.aspx?orgcode=00000000&orgtypecode=5&leftNavId=814&fycode=2011.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M. and Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability. The RAND Corporation. Retrieved on November 27, 2011 from http://www.rand.org/pubs/monographs/2004/RAND MG158.pdf.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M. and Hamilton, L. S. (2004). The promise and peril of using value-added modeling to measure teacher effectiveness. RAND Education Retrieved on November 27, 2011 from http://rand.org/pubs/research_briefs/2005/RAND RB9050.pdf.

Millar, A., Simeone, R.S., and Carnevale, J.T. (2001). Logic models: A systems tool for performance management. *Evaluation and Program Planning, 24*(1), 73-81.

Oakes, J (1986). Keeping track, part 1: The policy and practice of curriculum inequality. *Phi Delta Kappan, 68*(1), 12-17.

O'Malley, K., McClarty, K., Magda, T, & Burling, K. (2011). Making sense of the metrics: Student growth, value-added models, and teacher effectiveness. *Pearson Bulletin, 19*, 1-5.

Penuel, W., Fishman, B.J., Gallagher, L.P., Korbak, C., & Lopez-Prado, B (2008). Is alignment enough? Investigating the effects of state policies and professional development on science curriculum implementation. *Science Education*, 93(4), 656-677.

Petersen & Cruz (2004). What did we learn today? The importance of instructional alignment. *Strategies, 17*(5), 33-36.

Phillips, S. E. (2009). Legal corner: Using student test scores to evaluate teachers. *NCME Newsletter, 17(4)*, 3-6.

Piche, D. (2007). Basically a good model: NCLB can be fixed. *Education Next, 3*, 57-59.

Popham, W. J. (1978). The case for criterion-referenced measurements. *Educational Researcher, 7*(11), 3-10.

Porter, A.C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3-14.

Porter, A. C., & Smithson, J. L. (2001). *Defining, Developing, and Using Curriculum Indicators. CPRE Research Report Series* (No. RR-048). Philadelphia, PA: Consortium for Policy Research in Education.

Porter, A.C., Smithson, J., Blank, R., & Zeidner, T. (2007). Alignment as a teacher variable. *Applied Measurement in Education, 20*(1), 27-51.

Public Agenda and Education Week, Reality Check 2000. January 2000.

Remillard, J.T. & Bryans, M. B. (2004). Teachers' orientations toward mathematics curriculum materials: Implications for teacher learning. *Journal for research in mathematics Education, 35*(5), 352-388.

Scherff, L. & Piazza, C.L. (2008). Why now, more than ever, we need to talk about opportunity to learn. *Journal of Adolescent & Adult Literacy, 52*(4), 343-352.

Sherin, M. G. & Drake, C. (2010). Curriculum strategy framework: Investigating patterns in teachers' use of a reform-based elementary mathematics curriculum. *Journal of Curriculum Studies, 41*(4), 467-500.

Smith, C. (2008). Design-focused evaluation. *Assessment & Evaluation in Higher Education, 33*(6), 631-645.

Soto, A., Sireci, S.G., Keller, L.A., & O'Malley, K (2011). Evaluating Teachers Using Value- Added Models: Current Practices and Validity Evidence. *Center for Educational Assessment Research Report No. 792*. Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst.

SPSS, Inc. (2008). SPSS Statistics for Windows (Version 17.0) [Computer Software]. Chicago, IL: SPSS, Inc.

Steele, J. M. (1970). *Assessing intent and practice in instruction*. Paper presented at the 54th Annual Meeting of the American Educational Research Association and the National Council of Measurement in Education, Minneapolis, MN, March 2-6, 1970.

Stein, M. (2004). Mathematics curriculum evaluation framework. *Journal of Direct Instruction*, *4*(1), 41-52.

Tarr, J.E., Chavez, O., Reys, R.E., & Reys, B.J. (2006). From the written to the enacted curricula: the intermediary role of middle school mathematics teachers in shaping students' opportunity to learn. *School Science and Mathematics, 106*(4), 191-201.

Tennessee Department of Education (2011). Curriculum frameworks, maps, and pacing guides. Retrieved on December 11, 2011 from http://www.stemresources.com/index.php?option=com_content&view=article&id=130&template=map&Itemid=176.

Vaznis, J. (2011, June 29). Student scores to be a key factor in teacher evaluations. *Boston Globe*.

Wiggins, G. & McTighe, J. (2005). *Understanding by design*. Alexandria, VA: ASCD.

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

Yalow, E.S. & Popham, W.J. (1983). Content validity at the crossroads. *Educational Researcher*, *12*(8), 10-14, 21.